NOAA

Global Earth Observation Integrated Data Environment (GEO-IDE)

Concept of Operations

Version 3.3

13 September 2006

Prepared by:

U.S. Department of Commerce

National Oceanic and Atmospheric Administration (NOAA)

Data Management Committee (DMC)

Data Management Integration Team (DMIT)

Contents

Preface

In 1992, Congress ordered that NOAA biennially assess the adequacy of its environmental data and information systems. Of particular concern are the interfaces to these systems. NOAA data systems and those of other Federal agencies with environmental responsibilities should facilitate integration and interpretation of data from different sources. Partly in response to the latest assessment, and partly driven by the advantages offered by new capabilities in information technology, the NOAA Data Management Committee (DMC), in October 2004, called for development of an integrated NOAA data management plan. The NOAA Data Management Integration Team (DMIT) was convened to address this task. The membership of DMIT was selected to broadly represent data management activities and requirements within NOAA. The individuals on DMIT were selected based upon their experience, and technical insight. This document is the work of DMIT.

A separate document, the *GEO-IDE Implementation Plan* describes the actions, responsibilities and milestones needed to guide implementation of the GEO-IDE data integration strategy described in this document.

Executive Summary

To carry out its mission, NOAA must understand and address the complexity of many environmental problems and answer questions addressing contemporary societal needs. To do this, NOAA must be able to successfully integrate information from all of its goal areas and exchange data with partners in the national US-Global Earth Observation System (US-GEO) and the international Global Earth Observation System of Systems (GEOSS). Today, integration of data from multiplies sources or disciplines is difficult and expensive, which can lead to no or wrong answers to important societal questions.

With its Global Earth Observation Integrated Data Environment (GEO-IDE) as its contribution to US-GEO, NOAA will be able to provide easier and more cost-effective access to all of its data and information. NOAA will be better able to provide timely and accurate answers to important scientific questions and will better serve its customers; Federal, state and local governments, academia, the private sector and, ultimately, the American people.

NOAA's GEO-IDE is envisioned as a "system of systems" – a framework that provides effective and efficient integration of NOAA's many quasi-independent systems, which individually address diverse mandates in areas of resource management, weather forecasting, safe navigation, disaster response, and coastal mapping among others. NOAA Line Offices will retain a high level of independence in many of their data management decisions, encouraging innovation in pursuit of their missions, but will participate in a well-ordered, standards-based data and information infrastructure that will allow users to easily locate, acquire, integrate and utilize NOAA data and information.

The NOAA GEO-IDE will make NOAA products available in multiple formats and communication protocols, utilizing current information technology standards, where they are mature, and best practices, where accepted standards are still evolving. NOAA data and products will be described by comprehensive metadata that conforms to national and international standards. NOAA observing systems and collection, assimilation, quality control and modeling centers will provide their data and metadata in accordance with established NOAA GEO-IDE standards.

NOAA GEO-IDE will strive to take full advantage of the opportunities presented by internet technology to make access to environmental data and information as easy and effective as today's Web access to digital documents. It will also improve efficiency and reduce costs by bridging the barriers between existing, independent "stove pipe" systems and integrating the data management activities of all NOAA programs. It will do this through a federated approach, where the individual components retain a measure of responsibility and authority within the context of an overarching systematic set of goals, principles, and objectives.

Many NOAA information systems are critical to the national interest and we must ensure that improved integration and efficiency are achieved with minimal impact on these legacy systems and no interruption in essential services. Any changes in legacy systems that are needed for them to fully participate in GEO-IDE will be through the development of new interfaces to those systems and should not impact their basic capabilities.

GEO-IDE will fundamentally depend upon standards and it is essential that these be thorough, documented, and supported standards with demonstrated benefits. To ensure these standards are embraced and accepted across NOAA, an open and inclusive

“standards process” for nominating, evaluating, and implementing NOAA GEO-IDE standards is proposed. The standards process will define what standards are adopted, when they become effective, and how the organization will build up to and support the implementation of those standards.

To ensure standards are effectively applied across all of NOAA, project managers and developers must understand and use them. A NOAA Guide on Integrated Information Management will be developed to serve as a single reference point for NOAA data management polices and guidelines, an inventory of data systems, and relevant NOAA, national and international standards (in all stages of the NOAA approval process).

To achieve its goals, this plan recommends an incremental approach with continued operation of existing systems and standards while gradually improving integration through an evolutionary process of pilot projects and iterative improvement. Learning from and working with existing data integration initiatives, GEO-IDE will make use of standards, standard tools, and lessons learned. GEO-IDE aims to retain existing systems as much as possible while building a software infrastructure that links these systems together. This software infrastructure, called a Service Oriented Architecture, is a style of systems design based on using loosely coupled connections among independent programs to create scalable, extensible, interoperable, reliable, and secure systems.

Service-based architectures have been proven to solve interoperability problems including integrating systems developed in various programming languages, running on different computing environments and developed by autonomous groups at different times. These architectures make it practical to adapt and connect existing systems quickly for accomplishing new tasks and to benefit from highly evolved and still useful “legacy” applications.

Good governance is critical to the successful implementation of GEO-IDE. A higher level administrative structure that provides a suitable context for the Governance of GEO-IDE already exists: the NOAA Observing System Council and Data Management Committee (DMC). The DMC established the Data Management Integration Team (DMIT) to develop the GEO-IDE Concept of Operations and provide expertise and advice on the near-term (5-year) actions needed to implement this plan. DMIT includes representatives from all NOAA line offices and goals. To ensure synergy and effective coordination with IOOS DMAC activities, all NOAA members of the DMAC Steering Team are also members of DMIT.

Many of the governance issues will require a detailed understanding of information technology and an expanded structure is recommended to oversee implementation. A number of GEO-IDE implementation teams will be assembled to define the detailed architecture and coordinate development of specific Web services. These teams will be guided by a full-time project manager, hired to oversee implementation of GEO-IDE.

Realization of the GEO-IDE vision will take years. Implementation will be pursued through a number of concurrent activities, following a spiral, iterative development approach. A companion document, *The GEO-IDE Implementation Plan* defines specific actions, responsibilities, and milestones needed to implement GEO-IDE over the next ten years. It calls for implementation to begin with the following high-priority activities to:

1. Establish the GEO-IDE project management structure.
2. Secure funding to support GEO-IDE activities.
3. Identify major information management systems in NOAA.

4. Evaluate, adopt and adapt information management standards within NOAA and publicize them via an on-line NOAA Guide to Integrated Information Management.
5. Define a NOAA-wide, service-oriented Web architecture.
6. Test the feasibility of utilizing a “data typing” approach to NOAA data and refine the categorization of data types used throughout NOAA.
7. Develop/acquire technical knowledge and skills.
8. Identify technologies for implementation of the SOA, define core Web services needed and implement these services via pilot projects.
9. Investigate new technologies to support the NOAA mission.

1. Introduction

NOAA's mission is

> *"To understand and predict changes in the Earth's environment and conserve and manage coastal and marine resources to meet our Nation's economic, social, and environmental needs."*

To carry out this mission, NOAA must be able to successfully integrate information from all of its goal areas to understand and address the complexity of many environmental problems and answer questions that are important to address contemporary societal needs. Furthermore, NOAA must be able to exchange data with partners in the national US-Global Earth Observation System (US-GEO) and the international Global Earth Observation System of Systems (GEOSS). With the Global Earth Observation Integrated Data Environment (GEO-IDE) as its contribution to US-GEO, NOAA will provide easier and more cost-effective access to all of its data and information. NOAA will ensure its data and products are collected and managed in accordance with policies, procedures and standards that support and enhance integration and conform to NOAA Administrative Order (NAO) 212-15. These activities will ensure that society can access and use high quality, complete, and integrated information needed to support critical environmental and societal decisions.

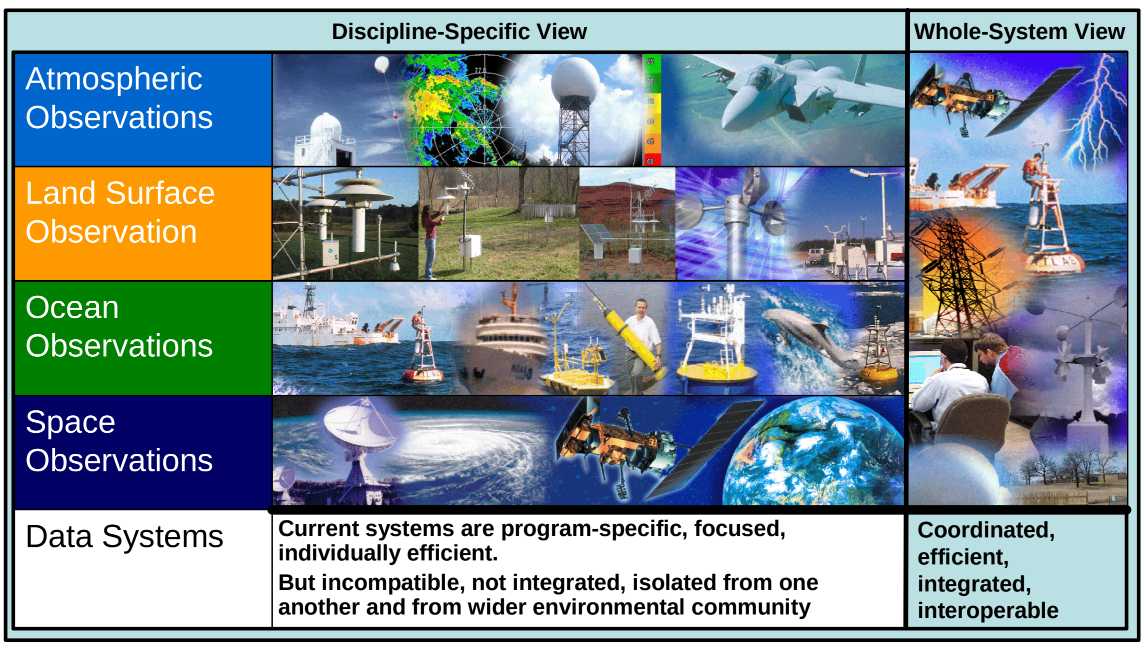

Figure 1.1 - An integrated, whole-system view is needed for coordinated and efficient operations

Over the past decade the advent of the Web and its attendant search engines has greatly improved access to documents and text that are available on line. However, this revolution in access to documents has highlighted how far we have to go to improve access to digital data. Web search engines cannot extract information from digital data holdings and no single standard guides the transfer of digital data and products over the Internet. Instead, as was true with documents before the Web, digital data are indexed and cataloged by many different sources and maintained and supplied in a multitude of formats. It is difficult and inefficient to locate data and hard to make effective use of data that are retrieved.

1.1. Goals

One goal of NOAA GEO-IDE is to take full advantage of the opportunities presented by internet technology to make access to environmental data and information as easy and effective as access to digital documents over the Web is today. Just as the Internet and Web browsers interoperate to make the location of documents nearly irrelevant, so should the process of locating datasets and individual elements from datasets be made effortless. Once located, analysis and visualization programs should be able to easily access, analyze and integrate data from many sources, regardless of their location or the underlying data storage techniques in use.

Another important goal is to improve efficiency and reduce costs by bridging the barriers between existing, independent "stove pipe" systems and integrating the data management activities of all NOAA programs, while avoiding a fully centralized approach. A federated approach, where the individual components retain a measure of responsibility and authority within the context of an overarching systematic set of goals, principles and objectives is likely more achievable and cost-effective. Many NOAA information systems are critical to the national interest and we must ensure that improved integration and efficiency are achieved with minimal impact on these legacy systems and no interruption in essential services.

To achieve these goals, this concept of operations recommends capitalizing on on-going data management initiatives and continued operation of existing systems and standards while gradually improving integration through an evolutionary process of pilot projects and iterative improvement. It aims to retain existing systems as much as possible while building a software infrastructure that links these systems together. This software infrastructure, called a Service- Oriented Architecture, is a style of systems design based on using loosely coupled connections among independent programs to create scalable, extensible, interoperable, reliable, and secure systems.

Through the GEO Integrated Data Environment NOAA will:

- Identify and address gaps in existing data management systems.
- Create interoperability across data types, disciplines, space and time scales, etc.
- Develop and adopt standards for data access protocols and data formats.
- Develop and adopt standards for terminology, units and quantity names.
- Improve integration of measurements, data, and products.
- Define a Data Management Architecture to integrate existing systems and provide a framework in which to meet needs of future data systems.
- Improve the efficiency of NOAA business by eliminating barriers to information access and reducing duplication through development and implementation of a Service-Oriented Architecture.
- Make it possible for the vision of US-GEO and GEOSS to succeed.

1.2. Benefits

The NOAA GEO Integrated Data Environment will enhance our ability to integrate observations and products, improve quality control, modeling and dissemination and standardize discovery and access to NOAA data and products. This will greatly expand the effectiveness of in-discipline areas (*e.g*. research, marine forecasts, storm forecasts, disaster planning, disaster management, etc.) as well as allow improved use of information to address multi-disciplinary societal issues. It will enable access to data and information across various NOAA goals, programs and observing systems in timely, scientifically valid, and user-friendly ways.

Information from a variety of societal theme areas must be successfully integrated to address the complexity of many environmental problems. Consider what is needed to understand the societal impacts of sea level change along our coasts. Information from diverse areas including weather, climate, disasters, water resources, ocean resources, and ecosystems, as illustrated in Table 1.1, must be successfully integrated to address this problem.

Table 1.1 - Examples of how sea level integrates across theme areas.

Theme Areas	Important Observables	Time-scales of interest
Disaster reduction	Hurricanes and Tsunamis	Multiple time scales
Human Health	Safety	Episodic
Climate	Sea ice extent & land ice/ocean heat content	Weekly to decadal/annual
Water Resources	Land water withdrawals/ Coastal water tables	Decadal/Annual
Weather	Storms (winds/waves) and Storm surges	Daily to weekly
Ocean Resources	Sea level & detailed coastal elevations	Annual/Decadal
Agriculture & Land-Use	Coastal relief & infrastructure	Century/Decadal
Ecosystems	Coastal flora and fauna	Annual to decadal

As another example, the measurement and analysis of drought has many time and space scale dependencies that affect all of the societal theme areas. In this example full integration would address common observing, data, and analysis needs as applied to every one of our theme areas. Table 1.2 provides some examples of the kinds of data and information that would need to be integrated to address drought across themes.

Table 1.2 - Examples of how drought integrates across the themes

Societal Benefit Areas	Important Observables	Time-scales of Interest
Human health	Water availability/quality	Daily to seasonal
Energy	Reservoir and lake water levels	Monthly
Climate	Boundary conditions	Weekly to decadal
Water resources	Ground water and lake levels/	Seasonal to decadal
Weather	Circulation, water vapor	Daily to weekly
Ocean resources	River flow	Monthly
Agriculture	Soil moisture	Weekly
Ecosystems	Water availability/quality	Weekly to decadal

Development and implementation of the Service-Oriented Architecture described in this document will improve the efficiency and effectiveness of data and information management systems within NOAA. This approach has a proven record of solving interoperability problems which include integrating systems developed in various programming languages, running in different environments on heterogeneous compute platforms, and developed by independent groups in autonomous organizational units at different times. It provides a means to improve integration and interoperability and can lead to a great increase in the reuse of software across NOAA.

NOAA's GEO-IDE will improve the application within NOAA of standards and best practices defined in related plans such as the Integrated Ocean Observing System

(IOOS) Data Management and Communications (DMAC) Plan and the *Integrated Earth Observation Data Management Plan*. This will lead to improved integration of information systems within NOAA and interoperability of NOAA systems with those of other government agencies and the wider commercial and academic communities. These improvements will, in turn, help NOAA make better use of holdings of external data and information in fulfilling its mission.

1.3. Why Now?

Congress, in U.S. Code Title 15, Section 1537 (1) and Section 1537 (2) ordered that at least biennially the Secretary of Commerce shall complete an assessment of the adequacy of the environmental data and information systems of NOAA. In conducting such an assessment, the Secretary shall take into consideration the need for (among others):

- The development of effective interfaces among the environmental data and information systems of NOAA and other appropriate departments and agencies.
- The integration and interpretation of data from different sources to produce information that can be used by decision makers in developing policies that effectively respond to national and global environmental concerns.

Improved integration of data management activities is critical to the success of US-GEO. As noted in the Interagency Working-group for Global Earth Observations (IWGEO) *Integrated Earth Observation System (IEOS) Draft Strategic Plan* (pgs 60-61)

> The U.S. needs a comprehensive and integrated data management and communications strategy to effectively integrate the wide variety of Earth observations across disciplines, institutions, and temporal and spatial scales.
>
> There are three urgent needs for data management:
>
> - New observation systems will lead to a 100-fold increase in Earth observation data.
> - Individual agencies' current data management systems are challenged to adequately process current data streams.
> - The U.S. Integrated Earth Observation System, linking the observations and users of multiple agencies, compounds these challenges.
>
> Data management is a necessary first step in achieving the synergistic benefits from the U.S. Integrated Earth Observation System.

Uncoordinated development leads to inefficiencies, incompatibilities, and duplication of effort. Increased efficiency is needed to handle the expected exponential increase in data volumes that will occur over the next decade. To cope with this unprecedented increase in the volume of data to be managed, NOAA must begin to improve coordination and integration of its data management activities now.

Several plans have recently been developed that include reference to the need for improving integration and interoperability of systems that manage Earth observation-related data. These include the:

- *IOOS Data Management and Communications Plan (DMAC);*
- *IWGEO Strategic Plan for the U.S. Integrated Earth Observation System;*
- *IWGEO Integrated Earth Observation System Data Management Plan;*

- *Strategic Direction for NOAA's Integrated Environmental Observing and Data Management Systems;*
- *Chief Financial Officer (CFO) Request: NOAA's Integrated Environmental Observation and Data Management Program; and*
- *NOAA's Environmental Data Management: Integrating the Pieces.*

The opportunities presented by improving interoperability between geospatial data have also been recognized by the educational and commercial sectors and in many areas industry and academia are leading the way. There are several initiatives now underway that address issues directly related to locating, sharing, use, and integration of NOAA data. Among the most significant are the activities of the Open Geospatial Consortium (OGC) and World Wide Web Consortium (W3C), the Federal Geographic Data Committee (FGDC), continued development and evolution of national and international metadata standards, the spreading adoption of the Open Project for a Network Data Access Protocol (OPeNDAP), and the development of the E-Gov Geospatial One-Stop Portal – an interagency geospatial data resource.

1.4. Risks

Continuing to develop systems in an uncoordinated manner will lead to further incompatibilities and will further isolate NOAA programs from each other and from the wider environmental community. This will increase the difficulty in integrating information between programs and hamper NOAA's ability to address important multi-disciplinary cross-goal societal issues, *e.g.*, coastal erosion, water resources, etc. The development and institutionalization of isolated islands of data systems that evolve independently may make future integration very expensive and isolate communities of users.

There are also risks in adopting a vision this ambitious. Key risks are:

- While the basic technologies are sound, there are risks in utilizing new technologies that have not been applied to NOAA data systems, its high volume of data, and the requirement to conform to NOAA security policies and to work harmoniously with current data systems and network architectures.

- Many attempts to build, apply and adhere to standards for data and metadata have failed due to a lack of uniform commitment to the process. The risk will be that there is insufficient management and financial support applied to standards necessary for GEO-IDE to be successful.

The likelihood of success in developing and implementing an integrated data environment can be increased by setting realistic goals, adopting current best practices for software engineering and project management, and by maintaining agility to make necessary mid-course corrections.

1.5. Present situation

Existing NOAA information systems have been developed to meet diverse sets of requirements. In general, these systems have been developed by individual programs to meet specific needs and are, thus, focused in their approach and efficient at what they do. The multiplicity of systems operated for different programs has, however, resulted in incompatibilities, inefficiencies, duplication of effort and higher overall costs for NOAA as a whole. Even with systems connected to the same network, incompatible protocols and interfaces are an effective barrier to interoperability.

As illustrated in Figure 1.2, a multitude of observing and data processing systems contribute data to support NOAA goals. Many of these systems are operated by NOAA, while others are operated by partner agencies that make their information available to NOAA or depend upon NOAA for long-term data archival. Data from these systems are encoded in many different formats and transmitted via a variety of communication systems and protocols. The amount, quality and format of metadata pertaining to these systems vary widely. Application of environmental data to multi-disciplinary problems is hampered by lack of agreed-upon and implemented standards needed to effectively identify, acquire, and correctly use all of the relevant data.

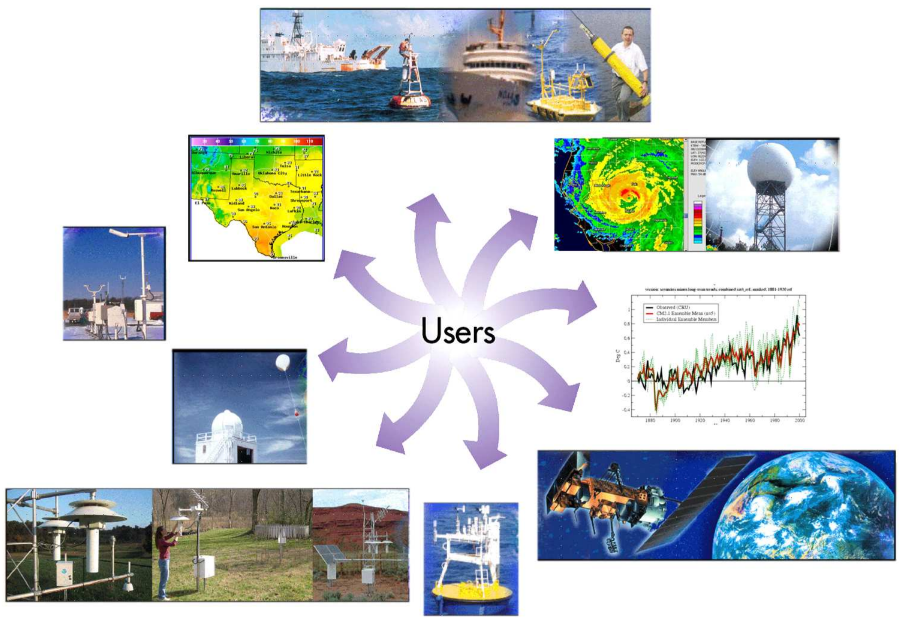

Figure 1.2 - Present situation: connectivity is limited and users must know where to access information. Data and products are available through incompatible interfaces and formats, and services from multiple centers cannot be easily combined.

1.6. Document organization

This document is organized as follows:

Chapter 1 *Introduction* - An overview of the goals, benefits and risks associated with the ideas presented in this document and other background material.

Chapter 2 *Scope* - The types of data, products and information management systems that are and are not covered by GEO-IDE.

Chapter 3 *Vision and Principles* - The vision for a NOAA Integrated Data Environment as well as a set of data management principles applicable to all NOAA environmental information systems.

Chapter 4 *Approach* - The technical and software development approaches recommended to implement the vision.

Chapter 5 *Governance Structure and Program Control* - An outline of the current organizational structure for oversight and coordination of NOAA-wide observations and data management activities. It also includes principles to guide program management efforts and decision making and a proposed organizational structure for implementation of GEO-IDE.

Chapter 6 *Towards a Service-Oriented Architecture* - Identification of specific items that are needed to achieve GEO-IDE.

Chapter 7 *The NOAA GEO-IDE Standards Process* - A proposed process for nomination, evaluation and implementation of information management standards for scientific and environmental data within NOAA.

Chapter 8 *NOAA Guide on Integrated Information Management* - An outline for an on-line NOAA guide, to help NOAA program and project managers implement information systems that conform to the GEO-IDE vision.

Chapter 9 *Priorities for action* - Priorities for action over the next 3 years (2007-2009).

2. Scope

This concept of operations defines a vision for applying a consistent set of principles, policies, and standards to the design, development, evolution, and operation of NOAA's data management systems. GEO-IDE shall facilitate convergence towards an integrated system that is aligned with NOAA's mission, goals and programs and is responsive to their requirements.

The NOAA Observing System Council (NOSC) has agreed that data management is defined by two coordinated activities: data management services and data stewardship. Together they constitute a comprehensive end-to-end process for movement of data and information from observing systems to data users. This process includes: data acquisition; quality control; validation; reprocessing; cataloging, documenting, storing and archiving the acquired data; and retrieving and disseminating the various data versions.

- **Data Management Services** include adherence to agreed-upon standards; ingesting data, developing collections, and creating products; maintaining data bases; ensuring permanent, secure archival; migrating services to emerging technologies; providing both user-friendly and machine-interoperable access; and assisting users and responding to user feedback.
- **Data Stewardship** consists of the application of rigorous analyses and oversight to ensure that data sets meet the needs of users. This includes documenting measurement practices and processing practices (metadata); providing feedback on observing system performance; validation of data sets; reprocessing (incorporate new data, apply new algorithms, perform bias corrections, integrate/blend data sets from different sources or observing systems); and recommending corrective action for errant or non-optimal operations.

Given the above definition, data management encompasses a wide range of information management functions, as shown below in Table 2.1. The boundaries between communication, data management and data processing systems can be ambiguous and subject to interpretation. Making optimal use of NOAA's data management systems for a variety of NOAA program requirements, while balancing the disparate, and sometimes contradictory, requirements placed upon them is a constant challenge.

Table 2.1 - Data Management Functions

Data acquisition	▪ Initial collection of raw data values ▪ Collection/creation of metadata ▪ Downlink and telemetry *are not covered by this GEO-IDE*
Ingest	▪ Transmission (Internet, private networks, satellite, media, etc.) ▪ Collection and storage of metadata ▪ Performance monitoring (observing, computing, communications, etc.)
Data Processing	▪ Data representation (format) ▪ Quality control (*e.g.*, detect missing data, check value limits, compare with neighbors) ▪ Quality assurance (*e.g.*, data validation, compliance with Data Quality Act)

	▪ Model/data intercomparison ▪ Aggregation in space and time ▪ Assimilation ▪ Modeling ▪ Production of products (charts, data records, warnings, forecasts, imagery, statistics, geodatabases, Internet mapping services, etc.) ▪ Analysis (means and extremes, trends, climate indicators, discontinuity and bias determination, statistical analyses, etc.) ▪ Reprocessing
Access	▪ Data discovery/catalogs ▪ Query - interactive browse or via intermediary personnel ▪ Data selection, extraction and translation ▪ Delivery of data, metadata and services (via telecommunications or media) ▪ Mapping and map services ▪ Visualization

As envisioned in the 2005 Report to Congress, an important focus of data management should be to ensure that NOAA data is easily shared within NOAA, with GEOSS participants and other user communities. This GEO-IDE concept of operations articulates the roles, methods, and standards to ensure that NOAA data are interoperable and easily transferred between these diverse communities of users. It establishes a process for identifying standards, policies and recommended tools to enable integration between independent systems that perform each of the data management functions identified in Table 2.1.

GEO-IDE focuses on the future state (*e.g.*, 5 to 10 years) of integrated NOAA data management. It provides the building blocks for a smooth evolution from the status quo to an integrated system of systems. GEO-IDE describes a framework for how on-going and new data management initiatives (*e.g.*, Comprehensive Large Array-data Stewardship System (CLASS), IOOS DMAC, Advanced Weather Information Processing System (AWIPS) modernization, etc.) should be developed to maximize data integration. Through GEO-IDE, NOAA will be able to identify, endorse or develop standards and protocols to effectively migrate legacy systems toward a common vision. GEO-IDE defines and prioritizes specific actions to pursue, and proposes responsibilities to implement integrated data management capabilities.

With respect to numerical modeling, the input and output of models are within the scope of GEO-IDE (*e.g*. data and file formats, communication systems and protocols, metadata, documentation and ultimate disposition of output products). While it is important to retain model source code, data inputs and outputs, etc., GEO-IDE does not address the way in which technical or scientific model algorithms are developed.

GEO-IDE is concerned with environmental and geospatial data and information obtained or generated from worldwide sources to support NOAA's mission. This is consistent with NAO 212-15, which includes the following definitions for these data.

> ***Environmental Data*** *- recorded observations and measurements of the physical, chemical, biological, geological, or geophysical properties or conditions of the oceans, atmosphere, space environment, sun, and solid earth, as well as correlative data and related documentation or metadata. Media, including voice recordings and photographs, may be included.*

> ***Geospatial Data*** *- information that identifies the geographic location and characteristics of natural or constructed features and boundaries on the Earth. This information may be derived from, among other things, remote sensing, mapping, and surveying technologies. Statistical data may be included in this definition at the discretion of the collecting agency.*

Thus GEO-IDE covers, for example, the following types of data, information and services:

- Chemical, geological, or geophysical properties or conditions of the oceans
- Chemical and physical properties of the atmosphere, space environment and sun
- Geological and geophysical properties of the solid earth
- Paleoclimatological and other proxy records
- Ecological and biological properties and conditions of the oceans
- Socio-economic data collected for or associated with a NOAA mission
- Information that identifies the geographic location and characteristics of natural or constructed features and boundaries on the Earth

GEO-IDE does NOT cover data or data management requirements for administrative support systems, such as finance, personnel, acquisition or facilities management.

3. Vision and principles

3.1. Vision

NOAA's GEO-IDE is envisioned as a "system of systems" – a framework that provides effective and efficient integration of NOAA's many quasi-independent systems, which individually address diverse mandates in areas of resource management, weather forecasting, safe navigation, disaster response, coastal mapping, etc. NOAA line organizations will retain a high level of independence in many of their data management decisions, encouraging innovation in pursuit of their missions, but will participate in a well-ordered, standards-based data and information infrastructure.

NOAA's GEO-IDE will provide friendly and flexible mechanisms to locate and access data and data products. It will address the needs of many classes of users including private industry, students and educators, researchers, government agencies, and the American public. It will also foster a community of private sector, value-added information product providers to address the needs of specialized groups.

GEO-IDE will make NOAA products available in multiple formats and communication protocols, utilizing current information technology standards where they are mature, and best practices where accepted standards are still evolving. NOAA data and products will be described by comprehensive metadata that conform to national and international standards. Descriptions of products and data will be available over the Internet and searchable via standardized data discovery portals. NOAA observing systems and collection, assimilation, quality control and modeling centers will provide their data and metadata in accordance with established NOAA GEO-IDE standards. When a user feels a need for personal assistance, the GEO-IDE Web portals will guide the user to contact points – email help desks and telephone-based guidance.

NOAA's GEO-IDE will be a component of US-GEO. It will provide users with integrated access to data and information from other systems within US-GEO, smoothly integrating across Federal Agency, public-private, and inter-disciplinary boundaries. GEO-IDE will contribute data into US-GEO in accordance with US-GEO and GEOSS data and information standards and protocols. The planning of GEO-IDE will emphasize a sustained, close collaboration with and leadership within US-GEO and GEOSS.

GEO-IDE will satisfy the diverse requirements of operations, research, monitoring, and archives. It will provide reliable discovery and delivery of data from measurement subsystems to operational modeling centers and to users and the delivery of computer-generated (model) information. For users who have time-critical requirements, collection, transmission, processing and delivery of this information will be done in real time. GEO-IDE will enable research systems to provide data to operational centers, when this is deemed appropriate. It will ensure that all appropriate data flow seamlessly into and out of secure, long-term archive facilities.

GEO-IDE will provide a continuous, vigorous outreach process to identify and remedy difficulties encountered by any users. Through its governance mechanisms GEO-IDE will assure a continual assessment of changing user requirements and emerging technological solutions. Continual innovation will be a hallmark of GEO-IDE.

To illustrate the benefits that GEO-IDE will offer to users, consider the following scenario:

Today, NOAA struggles to provide its environmental information to customers in a way that makes it easy to locate, acquire, and use. For example, if a customer wishes to study coastal erosion and its impact on estuarine ecosystems, several research websites within NOAA, including NWS weather Web pages, NESDIS/NCDC climate Web pages, and the NOS NowCoast must first be located. Requirements with customer service representatives from several organizations would probably need to be discussed, since there is no single comprehensive gateway to all NOAA data. Once the relevant information has been located, each NOAA organization would likely have a different process to follow in order to acquire the data. Once obtained, the data would likely be in inconsistent formats, using inconsistent parameter names, units, and quality control. The documentation available to describe each dataset would vary widely. These problems are exacerbated if the customer needs data and information related to several scientific disciplines, such as meteorology, oceanography and ecosystems.

Under NOAA's GEO-IDE the steps to address the customer needs described above will look radically different. The customer's favorite application or a standard Web browser will provide access to all NOAA (and related) data and information through a single interface. The interface provides intuitive tools to locate data that may be of interest, allowing for refined searches based upon geographic region, date, discipline, parameters of interest and a host of other descriptive information (metadata). The data discovery server responds to requests within moments, with an assurance that it has comprehensively searched NOAA's data holdings and has identified all information that matches the request. The customer is then able to read descriptions of the data and browse images and visualizations in order to quickly evaluate the data and arrive at the subset of interest. All of the desired data, products and information can then be obtained in a manner compatible with preferred analysis tools and using standard terms and units. There is no need for awareness of the physical location of the data or the manner in which it is managed – the data subsets that are of interest are delivered in a ready-to-use manner. Thus, all information can be easily combined and analyzed without regard to its source. Comprehensive information about the data (metadata) is available to aid in understanding the corrections, adjustments and other processing applied to the data.

Customers can also benefit from services that are not available today. For example, data subscription services and application-supported data discovery services would allow the use of relevant data without a customer having to explicitly discover and access it.

The customer is provided with information on how to contact a NOAA expert for additional help if any problems or questions arise.

The data environment to be created by NOAA GEO-IDE is outlined in Figure 3.1.

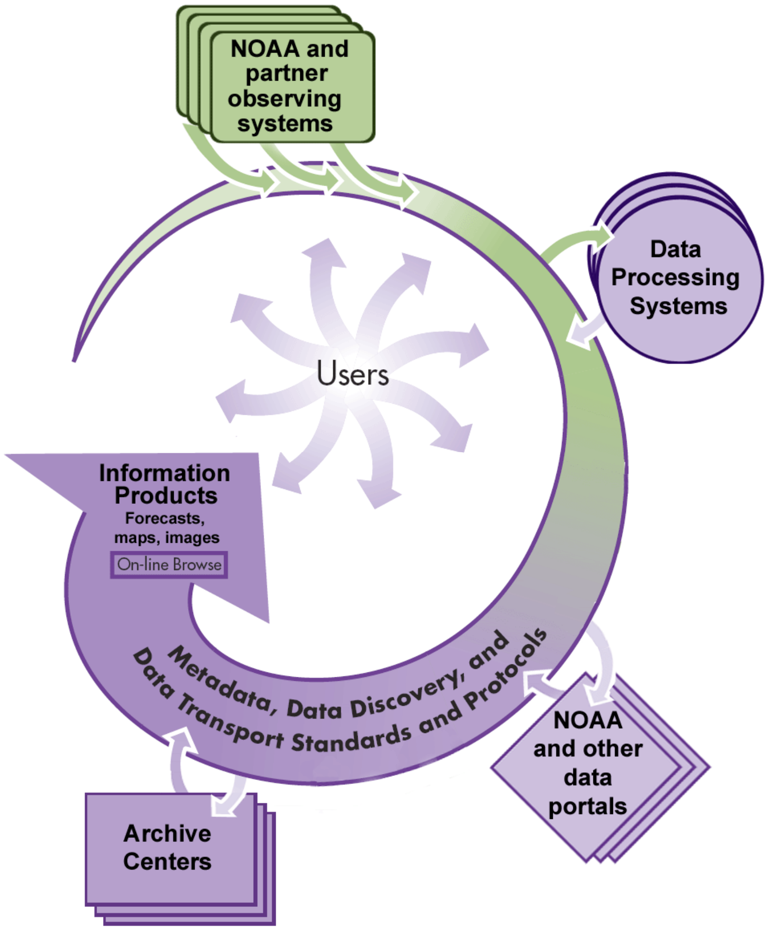

Figure 3.1 – Architectural vision

GEO-IDE depends fundamentally upon standards, and it is essential that these be thorough, documented, and supported standards with demonstrated benefits. To ensure these standards are embraced and accepted across NOAA, an open and inclusive "standards process" for nominating, evaluating, and implementing NOAA GEO-IDE standards will be adopted. The standards process will define what standards are adopted, when they become effective, and how the organization will build up to and support the implementation of those standards. The GEO-IDE governance infrastructure will assure that all parts of NOAA receive the training and support they need to successfully and usefully implement GEO-IDE standards.

3.2. Data Management Principles

Effective realization of this vision requires all NOAA data management systems to consistently follow a set of standard data management principles. Recommended principles are described below, including how they will be applied within NOAA:

1. ***Commitment and leadership:*** *Information is a strategic asset and information management must be a key component of every environmental data and information program. This ethic must be reflected in a corporate culture, embraced throughout the organization that recognizes data as a corporate resource.*

NOAA management will be visible advocates for development and implementation of NOAA-wide information management investments, policies, and procedures. All NOAA employees and contractors are stakeholders in the integrated information management vision, and will strive to help the organization develop and implement policies and practices for achieving it. NOAA will establish mechanisms for ongoing communication, coordination and training to ensure that all its data producers have the knowledge and resources needed to implement NOAA data management policies.

2. ***Stewardship:*** *People who take observations or produce data and information are stewards of these data, not owners. These data must be collected, produced, documented, transmitted and maintained with the accuracy, timeliness and reliability needed to meet the needs of <u>all</u> users.*

   NOAA will strive to meet the requirements of all users in planning, developing, and implementing its data management systems. NOAA will endeavor to make the most of every observation it takes and data product it produces.

3. ***Long-term preservation:*** *Irreplaceable observations, data products of lasting value, and associated metadata must be preserved. This information must be well-documented and maintained so that it is available to and independently understandable by users, now and in the future.*

   NOAA will ensure all data, products of enduring value, and associated metadata are well documented and maintained in suitable archives. NOAA, in concert with its users and partners, will establish criteria and procedures to guide the acquisition, documentation, retention, and purging of data to ensure important and irreplaceable information is maintained for posterity.

4. ***Requirements-driven:*** *It is essential that providers and users of data and products play an active role in defining the constantly evolving requirements that drive the development and evolution of data management systems.*

   NOAA understands that it has unrealized potential for the use of its data and information. NOAA will work with its growing and increasingly diverse set of data providers and users to determine present and future environmental requirements and applications and to continuously improve its relationship with both groups. NOAA will establish a vigorous outreach process to involve both groups and to help identify where improvements are needed. NOAA will foster development of a value-added "market", in which others may readily produce information products tailored to particular groups.

5. ***Discovery and access:*** *Freedom of access, mechanisms that facilitate discovery, timely delivery, use and interpretation of data and products (directories, browse capabilities, metadata, mapping, visualization, etc.) are essential (while following relevant policies and regulations).*

   NOAA will develop information systems and tools to facilitate discovery, use, and interpretation of data and products by its users. It will work with its partners in government, academia, and industry to make sure its data are available and accessible to all, while respecting any data confidentiality agreements. NOAA will ensure timely access to data and products necessary to support operational and research requirements.

6. ***Standards and practices:*** *Appropriate use of information technologies, widely shared standards, and integration approaches are vital to facilitate collection, management, discovery, dissemination, and access services for environmental data and products. This will ensure interoperability among providers, systems, and users. Effective application of standards and best practices contribute to the development of systems that are interoperable, efficient, reliable, scalable, and adaptable.*

   NOAA subscribes to the value of, and need for, corporate standards, but also recognizes the need for flexibility so that individual creativity in getting jobs done is enhanced by the use of standards. NOAA will define a process for standards adoption that is open and inclusive, and fosters buy-in by all stakeholders. Existing information technology and scientific standards will be favored. NOAA data and information will be consistent to the extent that implementation at each level, and across units, is compatible and mutually supportive.

7. ***Quality:*** *Data, products and information should be of a quality sufficient to meet the requirements of society and to support sound decision-making.*

   NOAA will strive, as a commonly understood, accepted, and supported goal, to bring quality information to people and processes inside and outside of NOAA. NOAA, together with partner agencies and institutions will strive to ensure its environmental information is of the highest possible quality within reasonable cost. The quality of NOAA data and products will be evaluated, fully characterized, and documented.

8. ***Cooperation and coordination****: Environmental and scientific data management is a task of global scope – a whole that should be much bigger than the sum of its parts. It is only by participating in a global community of integrated data management that each organization can realize the potential of its data to the betterment of humankind.*

   NOAA will actively participate in and commit to utilizing data management solutions that are compatible and interoperable with data systems utilized by international partners; by other US Agencies; by private sector data suppliers and users; by the research community; and by end users at all levels of US society.

9. ***Security:*** *Data, information, and products must be preserved and protected from unintended or malicious modification, unauthorized use, or inadvertent disclosure.*

   NOAA will ensure that its data management systems comply with all applicable federal security policies. It will ensure the integrity of its stored and transmitted, data and will protect data, networks and services from unauthorized use or attack.

4. Approach

4.1. Introduction

NOAA is a diverse organization with many quasi-independent information systems, which individually address mandates in areas of resource management, weather forecasting, and safe navigation, among many others. These systems are critical to the national interest and GEO-IDE must ensure that its goals of improved integration and efficiency are achieved with minimal adverse impact to the functioning of legacy systems and no interruption in essential services.

The direct approach to this problem would be to develop an entirely new NOAA-wide environmental information system and to replace existing systems wholesale. Once the new system was completed, after a period of parallel operations, legacy systems would be turned off and the new system would assume their functions. Given the diverse requirements of NOAA programs and the large number of existing systems, such an approach would be extremely difficult and costly, and the risk of failure would be unacceptably high.

The preferred approach is to capitalize on on-going data management initiatives (*e.g.* the IOOS DMAC; the Fisheries Information System; the NOAA National Operational Model Archive and Distribution System (NOMADS); Global Earth Observations System of Systems; etc.) and continued operation of existing systems and standards (*e.g.* AWIPS, CLASS, Family of Services, etc.) while gradually improving integration through an evolutionary process of pilot projects and iterative improvement. GEO-IDE will take advantage of useful and mature existing systems, while building a software infrastructure that links these and other new systems together into an integrated framework.

4.2. Proposed Model: Service-Oriented Architecture

As identified in Section 3, the vision for the GEO-IDE is one of cooperative integration. The goal of integration is to retain existing systems as much as possible while building a software infrastructure that links them together. The approach proposed to implement such a vision is through the development of a software infrastructure called a Service-Oriented Architecture (SOA). SOAs are a style of system-of-systems integration based on using loosely coupled connections among independent systems to create a scalable, extensible, interoperable, reliable, and secure framework. SOAs have been proven to solve interoperability problems which include integrating systems developed in various programming languages, running in different environments on heterogeneous computer platforms, and developed by independent groups in autonomous organizational units at different times.

SOAs are built on a software technology called web services (Figure 4.1). In the figure, a Service Provider is the system offering a service. A Discovery Service (sometimes known as a service broker) is a well-known repository for information about other services. The Service Requestor is the system requesting to discover and use a particular service. The Publish connector line indicates a provider registering its services. The Find connector line represents queries made to discover details (where, and how to communicate). The Interact line represents the communications between the requestor and provider needed to obtain the requested service.

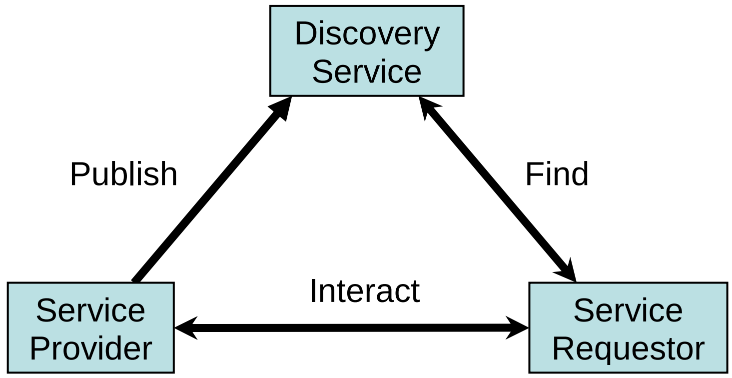

Figure 4.1 – The Web-Services Interaction Model

The standards body for the W3C defines a Web service as “a software system designed to support interoperable machine-to-machine interaction over a network.” Services insulate applications from the underlying platform (hardware and operating system) required to accomplish the task. Web services can be simple (authorization, searching, naming, registration), or complex, combining multiple services into a composite service that encompasses the comprehensive requirements of an application. For example, a user wishing to run a forecast model could utilize a complex service composed of services to: (1) build initial and boundary condition files on one machine, (2) send the results to a service on another system where the model is run, (3) generate forecast products using a third service and (4) visualize the results with a fourth service. Creating services to accomplish common tasks allows an organization to reduce the effort required to develop, port, and maintain its hardware and software systems. Composite services that make use of other services such as authentication and search reduce duplication of effort and increase software reuse, reliability, and security.

Service-based architectures address the two most important aspects of data management integration at NOAA: data sharing and application interoperability. Data sharing refers to standards and infrastructure to support sharing data that is stored in different formats, made available through different access methods, and provided by independent sources. The value of insights made possible by merging data from diverse sources within the same visualization or analysis justifies efforts to provide an infrastructure that supports data sharing across the NOAA enterprise and to external programs, organizations and users. Application interoperability refers to a framework that provides the ability for applications to communicate with and use Web services provided by other applications. A service-based infrastructure that allows independent programs to interoperate by communicating across a network will make it practical to build systems from reusable parts, to adapt and connect existing systems quickly for accomplishing new tasks, to benefit from highly evolved and still useful “legacy” applications, and to automate processes among different organizational units that currently require manual steps.

Building such an infrastructure from scratch is not necessary, since off-the-shelf and open source implementations of Web service infrastructure are available and will soon be included in most software development environments.

Two types of Web service standards are currently supported by industry: SOAP and REST. SOAP and the Web Services Definition Language (WSDL), combined, provide a way for service requestors and providers to exchange information through XML-formatted messages. These SOAP messages contain all the information needed to invoke a Web service through either a remote procedure called (RPC) or a Web service invocation. REST (Representational State Transfer) is a model for Web services based solely on HTTP. REST assumes that HTTP specifications provide all of the capabilities necessary for Web services and additional specifications, such as SOAP and WSDL are not required. Any item can be made available (*i.e.* represented) at a URI and, subject to the necessary permissions, it can be manipulated using one of the simple operations defined within HTTP (GET, PUT, POST, and DELETE).

In setting GEO-IDE standards for an SOA, it is important to recognize that these two approaches both have advantages and drawbacks, and there is no reason to standardize a single architectural style when different services may require different styles. In some cases, it may be necessary to support both ways of accessing a service, to make it integrate well with development tools and to provide a capability to evolve.

4.3. Web Services in NOAA

Web services can be described as a thin layer built on top of existing NOAA data management systems in which functional capabilities to access these systems are made available to the applications that require them. Additional Web services can be developed and added to GEO-IDE where functional gaps exist or new capabilities are required. Figure 4.2 illustrates a conceptual SOA for GEO-IDE (note the services listed are only representative; a comprehensive list is provided in Section 6.2). Conceptually, users access the information infrastructure to perform tasks that make use of information resources. The users' activities are supported by the fabric of the information infrastructure which may include shared hardware resources and long-term data archives. Users may be of many types: operational forecasting centers, state environmental management agencies, fisheries managers, individual researchers, etc.

The fabric of the infrastructure includes a set of components that support the use of the information resources. An important part of the fabric is one or more portals that provide the entrance point for users. Portals or Web-based graphical user interfaces permit users to locate and utilize distributed data or compute resources. One might envision data portals to access and utilize operational data, modeling portals to initialize and run weather or ocean models, and data management portals to monitor the state of NOAA's data systems. These portals could utilize common services including registries (to locate data sources), metadata systems (for information about data content), and ontologies (to map name spaces into a common language) to locate the appropriate Web services that meet their needs.

Information resources include datasets (*e.g.* observational data, processed data, model analyses, historical data collections), tools (*e.g.* quality control tools, analysis and visualization tools, open GIS software, software for generating derived datasets, event-detection software), numerical modeling modules (*e.g.* fully assimilative models to permit nowcasting, forecasting and data synthesis; model components that can be composed by users), and real-time data streams. Each resource is exposed within the organization as one or more services, *e.g.* as Web services, by which the resource is accessed or invoked. With this approach, only the way the service interface is described and accessed needs to be standardized, not the internals of the resource or the application in its local development environment. Thus, one "dataset" resource might be a NASA satellite image archive while another is a collection of ocean databases, some stored as flat files, some as relational databases etc, all accessible via data servers. Exposing the former as a service might just involve writing a Web service implementation of image search and retrieval while for the latter might involve setting up a server that runs a data access client that is essentially one or more Web services. Other services can be built to locate data through registries, monitoring and control services to insure critical systems are available, and services to insure the timely delivery of data through the appropriate operational or research network.

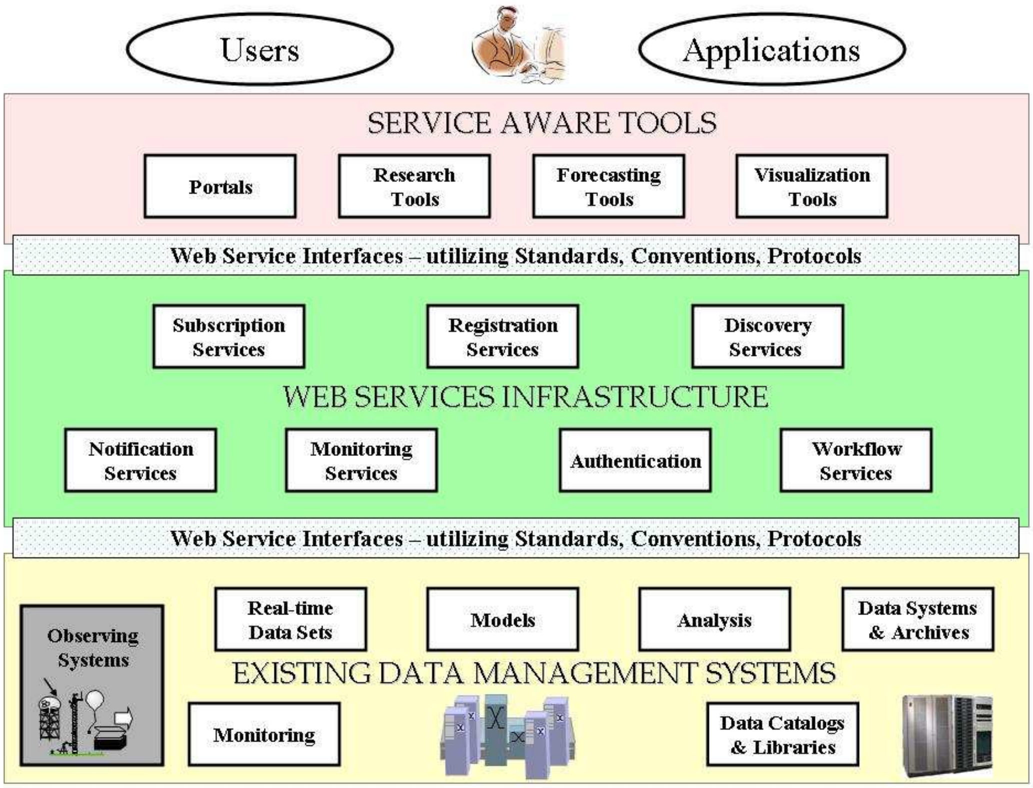

Figure 4.2 – A conceptual diagram of a Service-Oriented Architecture that integrates NOAA data management systems

Two classes of users must be considered: those whose identities are managed and those whose identities are unmanaged. Users with identities that are unmanaged, use a portal essentially like a public Web-page. While a portal may or may not ask such users to provide some information about themselves, it does not authenticate their identity, manage security certificates or provide other secure access for them, or maintain a personal workspace for them. Users whose identities are managed, first authenticate themselves with the portal to establish their identity; for example, using passwords, a secure electronic ID, or through a biometric identification procedure. For these users, the portal manages the users' security certificates that control access to resources within the organization.

An emerging technology that requires further investigation for its application to GEO-IDE is grid computing. Its goal is ubiquitous computing, where computing is available on-demand and users do not have to be concerned with where their tasks are running or where data reside. The most common analogy for grid computing is the electric power grid. Key to the success of the power grid has been the development and adherence to standards (*e.g.* voltage, amps, cycles, etc.). As GEO-IDE progresses through pilots to a distributed environment, developers monitor advances in grid computing and take advantage of tools and techniques that are developed.

4.4. Basis for development

Most of the primary standards required to build a service-based GEO-IDE are already available; NOAA will not need to create or define them. Web service standards are being adopted by the business and research communities as a means to build distributed interoperable systems. The two most widely used versions of Web services, SOAP and REST, are based on industry standards. Extensions to Web service standards are being developed to provide task and resource management functions across heterogeneous computing environments.

NOAA can also leverage existing distributed data technologies being used to link data providers with users via Services. For example, Open-source Project for a Network Data Access Protocol (OPeNDAP) servers have been deployed at numerous sites across NOAA to provide access to local, regional and global data sets on demand. These servers can provide information about the contents of model output or observational data, and can access and retrieve data for the requesting user or application. Other developments have been built on top of these services to provide added capabilities including servers to visualize model output, to handle new data formats and format conversions, and to build data catalogs. Several NOAA projects utilizing these capabilities include NOMADS to distribute model data, Meteorological Assimilation Data Ingest System (MADIS) to provide point data, and Live Access Server to handle oceanographic and other data. Demonstrated success in deploying distributed servers in heterogeneous environments has led to their being considered for use in the operational AWIPS system.

These developments represent a good basis for building GEO-IDE; however, significant work remains to define a common language (*e.g.* conventions, schemas, etc.) so communicating processes can understand each other using the underlying standards. For example, XML schemas must be defined so Web services can identify themselves in a common way to clients or other services requesting a service. Conventions will need to be established or adopted for time representation, parameter names, units, and data formats to facilitate information exchange among disparate distributed processes.

A management and architectural group, as described later in this document, will need to design and implement the SOA based on the analysis of current systems and future capabilities based on anticipated program goals and requirements. Technical committees must adopt, adapt, and if necessary, develop conventions and schemas that will be used to interpret requests and responses between communicating clients and Web services.

Four general classes of Web services are anticipated:

a. Operational Public Access Services: for public access to data, products and information services. Some examples include:
   - Electronic-commerce capabilities where required.
   - Subscription services so users can easily get the data they need when they need it. These could provide scheduled, event-triggered, or on-demand delivery mechanisms.
   - Common format translation.
   - Common coordinate transformation.
   - Visualization services.

b. Operational Services: where security, timeliness, and reliability are paramount. Some examples include:
   - Support for operational access to data (Warnings and Forecasts).
   - Subscription service.
   - Event notification service.
   - Format conversion.

c. Scientific Services: where efficient and flexible discovery and access to data sets are required. Some examples include:
   - Model initialization, invocation, and steering.
   - Access to local data (online), local offline (Mass Store/Archive Services), remote online (ftp, OPeNDAP, others), remote offline (remote-Mass Store/Archive Service, OPeNDAP, others).
   - Observing System Simulation Experiments.
   - Scientific Data Stewardship procedures and Archival Providence.

d. Commercial value-added services:

The responsibility of the design group will be to both identify and develop common services that satisfy needs from both the operational and scientific communities, and provide individual specialized services where programmatic or mission-specific requirements are demanded. Of course, the security of these systems will be addressed in their design.

A Notification and Data Subscription Service for Operations

A simple example of a Web service is subscription to a near real-time data stream of observations or model outputs. For example, an application for displaying regions susceptible to aircraft icing conditions might subscribe to a service providing meteorological parameters from observations and model outputs. The application would subscribe with a filter specifying the needed subsets of data and would include its own interface endpoint to which notifications would be sent. The notifications need not include the actual data, but merely a reference or query that could be used to access the data when available. Standards now exist for event-driven notifications as Web services, and off-the-shelf implementations of the necessary infrastructure are also available that provide scalable data subscription services to applications.

If such service interfaces were available for NOAA observational and model data, the current practice of polling an FTP directory every few seconds to see if desired data is available yet for download would no longer be necessary. Instead a much more scalable solution of event-driven notifications would provide timely access to applications that need real-time information for more complex processing.

4.5. Development Approach

NOAA GEO-IDE must encourage relatively small exploratory projects to build necessary services, one component at a time, between currently non-interoperable systems, to support the specific operational priorities described above. The results of such projects could quantify the levels of effort required to fully tie each part of the overall data infrastructure together. If successful, each such project would achieve a significant innovation and create an important foundation for further interoperability. However, it would be a mistake to evaluate an SOA approach by merely connecting two applications through Web service interfaces. Any such two-party connection can usually be provided with less effort directly, without the extra overhead of a SOA infrastructure. The real value of a SOA is its “network effect” that grows more rapidly than the number of services established.

The GEO-IDE data management and implementation processes must take place concurrently and in an on-going iterative, spiral development approach, where managers, architects, developers, and users work together. The implementation of GEO-IDE will most likely have a lengthy transition period while necessary services are developed and implemented. While the GEO-IDE architecture is being developed, initial core services can be advanced and provide building blocks upon which the architecture will grow as both requirements and technologies change. Local database managers and staff programmers must be provided the guidance necessary to begin to build, or modify existing applications to a more generalized loosely coupled solution. This way, the entire NOAA community will begin to build the system from the bottom up, but in accordance with NOAA-wide principles and standards.

Development of GEO-IDE will be based on an iterative, spiral-development process with the following stages:

- Select and evaluate pilot projects that relate to both operational and research parts of NOAA – especially those that show promise toward high levels of interoperability.
- Define standards, methods, schemas, security requirements, etc. necessary to interoperate within existing and emerging systems.
- Implement using standards, and demonstrate portability and interoperability of approach.
- Expand to new projects or capabilities and repeat the process.

4.6. Key Development Strategies

Both initial and long-term key development activities of NOAA GEO-IDE include the identification of pilot programs that employ a community-based open architecture design and that have adopted GEO-IDE guiding principles. These pilots will not only provide a jump-start for initial investment analysis, but provide a working set of prototypes.

Some key development strategies include:

- Building upon self-describing formats.
- Utilizing structural data typing to define, and classifying data and applications that require them.
- Determining initial and then follow-on services needed.
- Initiating pilot projects as recommended by the NOAA DMC that will advance or build the specific services discovering strengths and weaknesses of each.
- Following industry-and community-driven standards as appropriate.

4.6.1. Structural data typing

The GEO-IDE effort acknowledges that NOAA’s data systems are insufficiently integrated. This situation is a reflection of technology and management and decision-making strategies of the past that have tended to fragment data management, rather than to unify it. Lines of funding have traditionally been matched to observing system elements – satellites, ships, profilers, etc. – and data life cycle points – measurement, real-time applications, climate analysis, archive, etc. In the past, the observing system or function “owned” the data management specific to its system. Each observing system element has therefore developed individualized approaches to data management, often involving the development of unique (and non-interoperable) data formats and protocols. Real-time data management strategies were devised with little thought to analysis or archive, and so on. Predictably these traditions have hindered the development of integrated data management.

Communities of interest within data management are most naturally organized by structural type of data. The lines between these communities are drawn from the answers to key data management questions such as, what techniques are appropriate for searching for these data; for transporting (interchanging) these data; for visualizing or analyzing these data; and for storing or archiving these data?

Communities of interest defined by structural data types provide a natural way to organize data management efforts and specify standards required for interoperability. For example, the kinds of standards, best practices, metadata, and access interfaces required for time-series data collections are similar for atmospheric, oceanic, hydrological, biological, or climate data.

Traditional communities of interest defined by pattern of usage will continue to thrive of course, based upon scientific and societal goals. These communities will provide the requirements to an increasingly integrated data management community. For example, weather forecasters will continue to require synoptic access to observations; climate modelers will continue to view the same observations as time series. The role of the data management community will be to find unified solutions that address both of these usage patterns.

Table 4.1 proposes an initial list of communities of interest based upon structural data types. In most cases the structural data types are the natural consequences of the manner in which the data are collected. For any given data stream there may be ambiguities regarding the appropriate structural data type under which it should be handled. As a general rule, the best way to resolve this ambiguity is to choose the most highly ordered data type that could describe the data. Table 4.1 is presented roughly in order from most highly structured data types at the top to least structured types at the bottom.

Table 4.1 – Structural Data Types

Structural Data Class	Descriptions and subclasses	Examples and further explanation
Grids (and collections of grids)	• rectilinear grids • curvilinear grids • finite element meshes outputs • “unstructured” grids (variable numbers of vertices)	• finite difference model outputs • finite element model outputs • gridded (binned) data products • level 4 (gridded) satellite fields • spherical harmonic spectral coefficients[1]
Moving-sensor multidimensional fields (and collections of same)	• swaths • radials	• satellite passes • HF radar • side-scan sonar • weather radar

[1] In some cases, grids represent coordinate systems that are mathematically transformed from simple latitude-longitude-depth-time positions. Spherical harmonic spectral coefficients are an example of such.

Structural Data Class	**Descriptions and subclasses**	**Examples and further explanation**
Time series (and collections of time series[2])	• time-ordered sequence of records[2] associated with a point in space or a more complex spatial feature.	• ocean moored measurements[3] • fish landings at a port • stream flow records • sun spot activity • climate data (surface atmospheric stations) • paleorecords from cores, corals, tree rings, … • computed climate indices such as SOI
Profiles (and collections of profiles)	• height-or depth-ordered sequence of records[1] at a fixed (or approximately fixed) point in time and position in lat/long	• atmospheric soundings • ocean casts • profiling floats • acoustic Doppler instruments (structural overlap with time series)
Trajectories (and collections of trajectories)	• time-ordered sequence of records[2] along a path through space	• underway ship measurements • aircraft track data • ocean surface drifters • ocean AUV measurements
Geospatial Framework Data[4]	• lines • polygonal regions • map annotations	• shorelines • fault lines • marine boundaries • continually operating reference stations (CORS)
Point Data[5]	• scattered points	• tsunami or seismic occurrences • species sitings • geodetic control

[2] A “record” refers to one or more associated parameter values and associated metadata.
[3] Standards for time series need to consider small, time-dependent excursions in latitude, longitude, and depth. Cabled ocean moorings are an example of such.
[4] The “GIS perspective” must be a major focus in the discussion of all of the data classes listed in this table.
[5] As an organizing principle for data, “Point Data” is the lowest common denominator. Most structural data types are reducible to collections of points, though with a loss of essential semantics in most cases. For example, a grid may be represented as a collection of ordered tuples. Some types of measurements, for example tsunami occurrences or species sitings, naturally possess limited structure. For these measurements, the Point Data structure is the natural classification.
Note that real-time delivery of data will generally remove time structure, so that, for example, a collection of time series may reduce to Point Data when accessed in real time.

Structural Data Class	Descriptions and subclasses	Examples and further explanation
Metadata	• "data about data" – context information needed for the interpretation of data	Like other data types, metadata has distinct requirements for storage, access, archival, and transport. Metadata content is a major focus of discussions within all of the data types. Metadata as a "data type" refers specifically to its unique requirement and properties with respect to archival, access, and transport.

4.6.2. Advancing integration through pilot projects

Pilot projects serve as a means to both evaluate and identify weaknesses in current technologies, and to begin the process of building and integrating NOAA's data management systems. A test bed for data access and use is a fundamental building block for the development and implementation of many NOAA services. NOAA's SOA must be implemented with both legacy and emerging systems. Pilot projects to address these needs are required. Services solutions however must be generic in that they be general enough to accommodate both existing standards and emerging standards.

The OPeNDAP data transport interface is recommended to be used to provide a flexible basis for moving data between providers and users where operations like sub-setting, merging, formatting, and distributed data access are permitted. Adopting an OPeNDAP solution allows the possibility of reducing workload when integrating a new data source, or interacting with a new institution. The OPeNDAP technology is flexible and permits each institution to work using their favorite format or a basket of formats internally, but still maintains the goals of low-cost interaction with other institutions and ease of use. OPeNDAP in itself does not solve all data access issues, as application-specific knowledge of semantic structure and metadata layers remain. Semantic structures and metadata compatibilities will require convergence to naming schemas (*e.g.* the Climate and Forecast CF convention). Other services as outlined in Chapter 6 "Toward a SOA" must be built on a step-by-step basis with each new service adding services to the overall architecture. Key to the success of pull technologies is demonstration of host side data manipulation and sub-setting so that only needed data, not entire files, could be retrieved. Thus the data transfer time and network bandwidth requirements could be minimized. Agencies and institutions benefit greatly from an emphasis on low cost of buy in; *e.g.* keeping standards and protocols and software components simple and lightweight enough to be adapted and deployed without a dedicated team of local information technology experts.

Within NOAA's SOA, components need to be evaluated and merged where individual components provide one or more services to other services or to clients. These collections of components crossing NOAA goals will continuously adapt from user requirements in an iterative spiral developmental software engineering approach.

There are several areas where pilot projects will enhance NOAA's understanding of existing standards, permit NOAA to evaluate technologies which can be applied to data systems integration, speed the development of a systems architecture, and enhance the prospects for success of GEO-IDE. A list of perceived challenges and choices are given along with existing technologies that could be investigated and applied toward GEO-IDE. These include:

- Security: Explore security implications of Web services and methods to access proprietary data.
- Metadata: Apply proposed standards to NOAA data in order to identify and locate data.
- OGC: Investigate mechanisms to integrate OGC standards into NOAA data management systems.
- Data Transport: Explore data transport mechanisms to improve the movement of data across the network.
- Structural Data Typing: Categorize and build common mechanisms to access NOAA data for a specific community (e.g. Ocean Datasets). Extend to other communities when appropriate.
- Integration: Link CLASS and NOMADS under a common Web services infrastructure to support the discovery, access, and transport of data.

Recommended projects for application-specific use of semantic structure, client and server-side processing, standards advancement, and metadata resources include:

- GrADS Data Server (and its underlying “Anagram Server”)
- Live Access Server
- OGC Standards: Catalog, Web Services (OWS)[6], Coverage, Map, and Feature Services
- Earth Observing System Clearinghouse
- Global Change Master Directory (GCMD) Catalog Service
- Earth Observing Clearing House (ECHO)
- Open Abstract Data Distribution Environment (ADDE)
- THREDDS Data Server/Catalog Services

[6] For more information on OWS see http://portal.opengeospatial.org/files/?artifact_id=10380

5. Governance Structure and Program Control

5.1. Background

Effective management of NOAA's Data Management systems requires a strong governance structure and a well-defined process to ensure each program component is effectively monitored and appropriately managed. NOAA's initiatives in the integration program, given the magnitude and breadth of the data management program, require many components operating in an integrated and synchronized manner. A well-defined governance process and structure will better ensure that planning and control processes are constructed, resources are used wisely, and measurable results will be delivered.

Proper governance processes are crucial to how a program is managed. It will ensure that roles and responsibilities of all associated entities are clearly articulated; that the program is managed as a portfolio of projects, carefully selected according to clear, repeatable processes and objective criteria; that projects are well designed, properly implemented, and effectively managed; and the overall program performance is regularly assessed and evaluated.

A well-executed governance process will help protect the data management program from being distracted from achieving program goals and objectives. Having a sound, proven process for managing the performance and outcomes associated with this program is the best insurance against these pressures. A well-defined governance process will ensure we provide clear assurances that technology investments are necessary, purposeful, and will result in demonstrated improvements in mission effectiveness and serve society's needs.

The NOAA Data Management Committee (DMC) was established by the NOSC to coordinate the development and implementation of data management policy across NOAA. The DMC addresses issues and opportunities that require coordination among the Goal Teams, Line Offices, and Data Centers to address data management responsibilities. The DMC's objective is to provide clear guidance to NOAA on matters of data management and to provide the NOSC with the information it needs to bring about integrated data management within the NOAA Observing Systems Architecture. The DMC established the Data Management Integration Team (DMIT) to develop the GEO-IDE Concept of Operations and provide expertise and advice on the near-term (5-year) actions needed for implementation. DMIT includes representatives from all NOAA line offices and goals. To ensure synergy and effective coordination with IOOS DMAC activities, all NOAA members of the DMAC Steering Team are also members of DMIT. The relationships between these groups are illustrated in Figure 5.1.

With respect to information management standards, the DMC recommends that GEO-IDE be identified as <u>the</u> NOAA coordination group for all NOAA interactions with external standards activities relating to scientific, geospatial or environmental data and information.

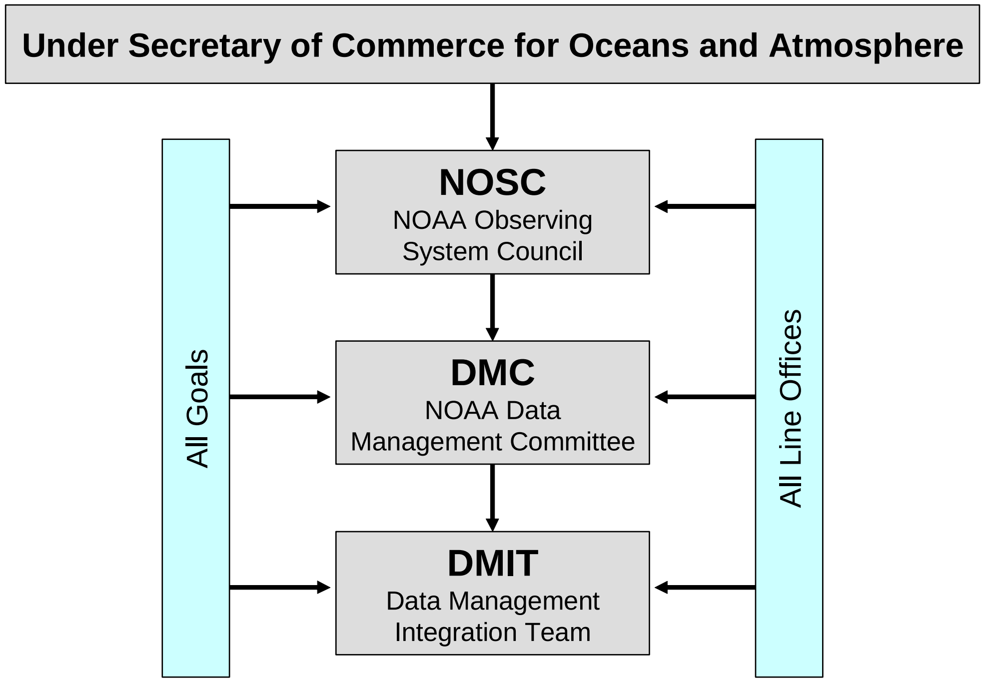

Figure 5.1 – GEO-IDE governance structure

5.2. Governance

5.2.1. Guiding principles

The success of GEO-IDE depends on properly run and coordinated operations at the global level. If the program is to be successful, effective management is required at all levels. Since GEO-IDE is a global program that aims to produce consistent and comparable data sharing and standards for all NOAA Line Offices, we must establish standards, and provide guidance to all levels of the organization. A governance process, by definition, requires discipline, consistency, collaboration, and communication. Provided below, are the principles that shall be used to guide program management efforts, management decision-making, evaluation, and pursuit of meaningful results:

Line offices and goal team participation – DMIT strongly encourages the full participation in the program from each line office and goal team. Business and technical expertise should be represented in all levels of the governance structure.

Transparency – All participants have a clear view into the governance process, program plans, project plans, business processes, and other elements and components of the program. Standard policies and procedures must be understood throughout all levels.

Proper representation – DMIT members represent both the program's national needs and their own agency. DMIT members need to have a broad understanding of NOAA's data management needs and also be sensitive to customized needs to satisfy practical situations and true business needs of specific programs. Committee members will wear a "big hat" when working on issues of national importance; "small hats" are appropriate when it is necessary to interpret the voice of the customer and to translate this into investment strategies that are relevant to all players.

Best practices – GEO-IDE will use industry best practices in program management, project management, and performance management. It is also important to understand that the program management/governance support activities will likely be significantly greater than that supporting late stage or mature programs.

Communication – In the early stage of GEO-IDE, a preliminary communications plan should be developed. A communication plan may evolve and be modified during program execution as changes occur. However, the communication framework should not be changed. Having a sound communication plan during the program implementation will ensure the transparency of the program's executive decisions and plans, the understanding of the policies and procedures, and a clear accountability.

Management science – While the overwhelming numbers of individuals that will interact and participate in GEO-IDE are highly educated and experienced scientists, biologists, statisticians, and information technology architecture and data management professionals, recognition of the contribution of the disciplines of management science and organizational development may have an impact on overall program performance and success.

Learning organization – A culture of capturing what is learned and carrying that forward as efficiently as possible to impact future efforts and initiatives; also includes knowledge-management techniques and fostering community best practices.

5.2.2. Stakeholders

Clearly identifying stakeholders of the program is another important success factor. There are numerous individuals and organizations that are affected by this program. At a high level it involves all of those that take an interest in any NOAA program or product, and also the individuals who are directly involved in data capture, analysis, and dissemination.

The diverse list of stakeholders includes:

- Congress
- NOAA
  - Senior NOAA Management
  - NOAA Program Offices
  - Data Management Committee (DMC)
- Other Federal Agencies and Organizations (e.g. USCG, State Department, DOI, NASA, DOD, IOOS, US-GEO)
- Other university, state and non-governmental organizations
- The private sector
- Emergency managers
- General public

5.2.3. Requirements and structure

5.2.3.1. Governance requirements for the GEO-IDE

Good governance is critical to the successful implementation of GEO-IDE. A higher level administrative structure exists (see Figure 5.1) that provides a suitable context for the governance of GEO-IDE. However, many of the governance issues will require a detailed understanding of information technology and an expanded structure is needed. To be successful, this expanded governance structure must embody the core functions of the DMIT (or group established to oversee GEO-IDE) and be able to effectively carry out the tasks described in the *GEO-IDE Implementation Plan*. Under GEO-IDE, the following activities need to be governed.

5.2.3.2. Functional area responsibilities

Data Management – This functional area will likely be the largest responsibility of the DMIT. Core data management functions have been identified by the DMIT (see Introduction). It will be the responsibility of this program to ensure that data management policy and standards are being adequately addressed for NOAA data management functions.

Standards – This functional area will manage the NOAA standards adoption process (see Standards section).

Service-Oriented Architecture – This functional area will oversee the development and implementation of the SOA plan for NOAA.

Program Operating Plans (POPs) and Planning, Programming, Budgeting, and. Execution System (PPBES) Process – This team will evaluate POPs and proposed systems to ensure that integrated data management is being adequately addressed, and be the liaison to the PPBES process.

Skills Development and Training – Once major data types and existing systems, standards, and best practices for managing those data types have been identified, NOAA data managers will need to acquire technical knowledge and skills for integrating their data into systems built using accepted standards and best practices. This migration will require significant training and support from internal and external experts. The DMIT Skills Development and Training Team will lead this effort.

Communications – Development of a communication strategy and plan is crucial to the success of this complex, large-scale, multi-location and agency program. Guidelines for the communication plan are provided in the *GEO-IDE Implementation Plan*.

Internal and External Coordination – To accomplish its goals, DMIT must coordinate with other committees, working groups, and offices both within and external to NOAA. This will be a functional area within and across DMIT.

A general assumption is that the existing NOAA management structure will provide the financial and in-kind resources needed to implement GEO-IDE and assign DMIT to lead the effort.

5.2.3.3. Operational requirements of DMIT

To carry out the activities listed above, it is recommended that DMIT remain in existence and a full-time project manager, the Data Management Integration Architect, be hired. Once the vision for GEO-IDE is accepted by the DMC, membership of DMIT should be re-evaluated to ensure that expertise and representation are sufficient to carry its implementation. DMIT should be comprised of highly qualified individuals from across NOAA who have experience and understanding of NOAA data types and information technology, but can also lead the functional requirements of the GEO-IDE. Annual operating plans will be developed at the beginning of each fiscal year and submitted to the DMC for approval. The basis for the annual operating plans will be the activities defined in the *GEO-IDE Implementation Plan*. Time commitments for the individuals selected to serve on the DMIT must be approved by the member's first-line supervisor and should be set at a realistic level. Ultimately, funding and DMC input will determine the size and staff allocation of the DMIT.

Coordination with related activities within NOAA

DMIT must coordinate with other committees, working groups, and offices within NOAA to accomplish its goals. Existing groups such as the NOAA Chief Information Officer (CIO) and line office CIOs, the NOAA GIS Committee, the NOAA Metadata Working Group, the NOSA Working Group, and others will provide valuable expertise and linkages to user communities/stakeholders within NOAA.

5.2.3.4. Proposed organizational structure

The following represents a proposed structure to support implementation of the GEO-IDE vision. It should be noted that this is a flexible structure likely to be modified as GEO-IDE matures.

The GEO-IDE Project Manager/Data Management Integration Architect and Business Manger/Administrator are at the top level of the DMIT and will coordinate DMIT functional and program areas (core activities). The proposed functional areas correspond to the GEO-IDE program components. These functional areas will be supported by DMIT members and by implementation teams to be established, which will undertake GEO-IDE's mission.

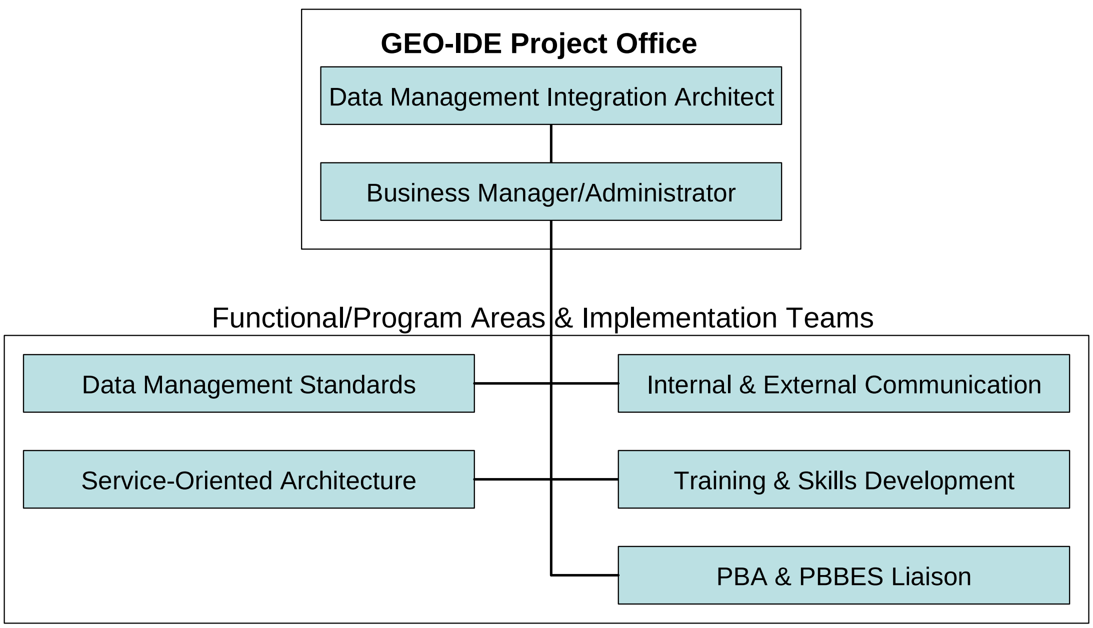

Figure 5.2 – Proposed GEO-IDE implementation structure

5.3. Management Components

5.3.1. Policy and procedure development

In order to carry out specific tasks and activities defined by the *GEO-IDE Implementation Plan*, DMIT needs to identify or develop program management policies and processes to optimize the outcomes of the program. These policies include a top-down communication structure, clearly identified program roles and responsibilities, program implementation and integration procedures, program working group policies, program standards and architecture, regulations, program control and evaluation, and rules concerning conflicts of interest, etc.

5.3.2. Management control

Section 5.2.3.2 lists functional governance areas of the program. Each of the functional areas depends upon governance to ensure success. The following sections describe general governance processes. Further development or refinement of these processes may be necessary as specific requirements of each functional area are developed.

5.3.3. Performance monitoring and management

Performance management is an integral element of enterprise transformation and program management. It defines how activities and components of the program will be measured and any shortcomings acted upon. Performance measures evaluate a program over time and are used to manage, track, and report progress, and to provide feedback for continuous process improvement.

GEO-IDE will use a performance measurement methodology to ensure the success of each program's activities and components. Each functional program area will need to follow performance management guidelines to develop a performance management plan and performance matrix to measure program progress against goals. Therefore, performance management will produce outcome-focused, results-oriented measures that can assist leadership in tracking results. Measures will need to change as progress is made. Measurement criteria should be defined as focused, appropriate, balanced, robust, integrated, and cost-effective.

5.3.4. Risk assessment and management

One of the governance activities involves risk management in order to mitigate risks and plan for contingencies. The purpose of risk management is to ensure levels of risk and uncertainty are properly managed so that the project is successfully completed. It enables those involved to identify possible risks, the manner in which they can be contained, and the likely cost of countermeasures and contingency plans for cases where predicted risks are realized.

Successful management of GEO-IDE requires informed, proactive, and timely management of risks. The emphasis is on cross-specialty, cross-discipline, cross-functional, and cross-technology development programs since GEO-IDE involves multiple organizations across NOAA. Such programs maximize risk opportunities and occurrences. The goal of risk management is designed to proactively identify and address risks early in the program and throughout the program life cycle in order to prepare for the unexpected and to adjust the program plan as needed. Risk management involves identifying, analyzing, controlling, and reporting risk factors throughout the program and measuring it against program objectives.

It is the responsibility of all team members to track risks and develop contingency plans to address risks. This falls within the role of the program management and governance components.

Risk identification begins in the early planning stages of a program. As scheduling, budgeting, and resource planning begins to evolve, the Risk Management Plan will change to reflect new risks identified in the planning stages and through the development stage. As projects progress, new risks may be added or removed based on changes during the various projects.

5.3.5. Project selection process

As noted earlier, this plan recommends capitalizing on on-going data management initiatives and continued operation of existing systems and standards while gradually improving integration through an evolutionary process of pilot projects and iterative improvement. DMIT will select pilot projects to be supported. The program governance will operate in the context of the “select, control, evaluate” framework commonly associated with capital planning and investment control in full lifecycle programs. This framework will help program decision-making teams select and finance the “right” portfolio of investments. Once selected, the governance process institutes project management controls, and an evaluation process will ensure that a funded project achieved its intended objectives within cost, schedule, technical, and performance baselines.

6. Towards a Service-Oriented Architecture

6.1. Data access and use

All NOAA data and products go through a similar series of steps between generation, processing, archive, and use. The actual divisions in this flow can be defined in myriad ways. Not all steps may be followed in any particular application and in many cases one or more of the steps will be invisible to a user. However, the overall chain of events is universal. The various stages in the process will be described in detail below, since integration must be supported within and between each step.

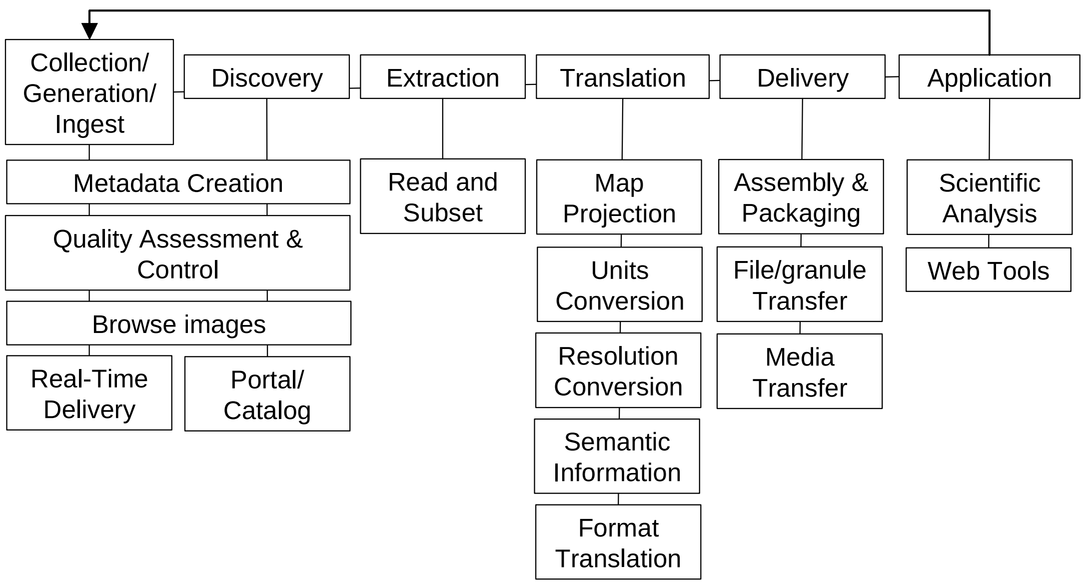

Figure 6.1 – Stages in data discovery, access and use

6.1.1. Generation / Ingest

Each of the stages outlined above has several requirements and software systems for addressing them. Some steps are executed in different ways during more than one phase. The first phase for every product is generation. This phase could include algorithms for analysis of satellite data streams or it could include field sampling by NOAA or other scientists. In some cases, other products are included in the generation process as ancillary datasets. Traditionally, the producers or collectors document data so that colleagues can use them, perform various quality control and assessment steps, may visualize the data and compare it with other datasets, and may deliver the data to users that can understand the metadata created by the producer and need the data quickly. This is the short path followed by many operational NOAA datasets and products today. To improve integration and ensure data and products are useful to the widest possible range of customers, the requirements of all potential users of this information must be recognized and addressed by NOAA. Metadata and other documentation must be extended with broader audiences and integration needs in mind.

After the data/products are generated they are ingested into operational processing systems and disseminated to real-time users. They are also transferred to archives for further quality assessment and dissemination to additional users. During ingest, additional metadata is usually created for long-term stewardship.

To improve integration of data and product generation and ingest, standards are needed in several areas:

1) Data/product representation (format) standards
2) Comprehensive metadata and documentation content standards

Modern standard formats for scientific data, such as HDF or NetCDF, combine data, metadata for those data, and access into a single interface. Using standard formats for all data generated by NOAA would significantly decrease resources required for data management integration and greatly simplify integration of multiple NOAA datasets. Therefore, the following steps are needed:

- Examine products for compatibility with standard formats and submit suitable formats to the standards approval process.
- Once standard formats are agreed upon, generate products using those formats.

To ensure that maximum value can be obtained from NOAA data and products, it is essential that comprehensive metadata and documentation be provided. This must be sufficient for both specialists and non specialists and include information on how and where the data were obtained. This allows the user to evaluate the quality of the data and determine if the data or products are applicable to their specific requirements. Figure 6.2 illustrates how different types of metadata are generated at different points in the data life cycle. Data generators provide data directly to operational users. These users need metadata that helps them use the data in routine processing. Much of this metadata changes with each file and is termed use-level metadata. Archive users, on the other hand, need metadata that helps them discover datasets and understand them without needing to interact with the data generators. This *discovery-level* metadata tends to change slowly and can be considered to be *quasi-static*. Metadata standards help ensure that these users can access and understand metadata from the archive.

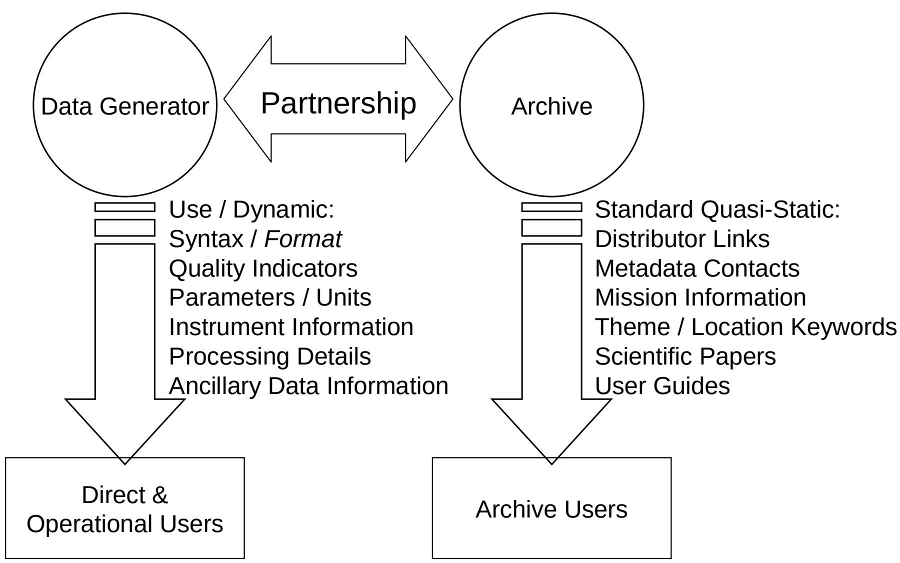

Figure 6.2 – Metadata sources, content and users

Existing metadata standards (Content Standard for Digital Geospatial Metadata (CSDGM) and ISO 19115) provide a good starting point for defining comprehensive use-level metadata. However, given the wide range of data and products created and managed by NOAA programs, these standards are probably not sufficient to meet all NOAA needs for documentation and metadata. Therefore, CSDGM and ISO 19115 must be evaluated to determine if extensions or additional elements will be needed to comprehensively describe all types of data managed by NOAA (especially ecological data).

In addition, as an on-going effort to improve the information and usefulness of data in archives, ISO has encouraged the development of standards in support of the long-term preservation of digital information obtained from observations of the terrestrial and space environments. ISO requested that the Consultative Committee for Space Data Systems (CCSDS) Panel 2 coordinate the development of those standards. (CCSDS has subsequently reorganized, and the work is now situated in the Data Archive Ingest (DAI) Working Group.) The initial effort has been the development of a *Reference Model for an Open Archival Information System (OAIS)*. The OAIS Reference Model has been reviewed and in its final stages of approval as an ISO Standard and as a CCSDS recommendation. GEO-IDE fully supports the implementation of OAIS by all archive centers and facilities within NOAA.

6.1.2. Discovery

Before a scientist or decision maker can use scientific data or products they must first be aware that the information exists. In most cases, people are aware of and familiar with information they need on a regular basis. However, this is not the case for data that could be of value but are from sources that are outside their normal experience. In this case, the user must search for and “discover” data or products that might be of interest. This normally involves consultation with colleagues, searches of data catalogs, examination of documentation and, possibly, reviews of browse images that provide a visual summary of the data.

Data discovery has traditionally been a barrier to effective utilization of data. In the past, discussions among colleagues within scientific communities were the primary source for exchange of information on the availability of data. As environmental science has become more interdisciplinary, it has become increasingly important for people to be able to discover data that could potentially be of value, especially from sources outside a user’s normal scientific community.

With the expansion of the Web, search engines have become an increasingly important tool for locating information. They have been very successful in locating documents and text, primarily because widely accepted standards govern the way text is defined on the Web and, most importantly, words have agreed meanings (within one language anyway). However, generic search engines have not been effective at locating scientific data because numbers are only meaningful if they are explained and documented. To make these data as easy to find as texts are today, this documentation, or metadata, must be written in accordance with well-defined and widely accepted standards.

Metadata refers to a wide range of information that describes data. At the highest level, discovery level metadata describes data collections in general terms. As the name implies, this provides information to help a user discover if data of interest exist and where they might be obtained.

Locating the data of interest is only the first step in deciding whether or not those data will be relevant to the proposed objectives. A comprehensive, in-depth description of the data specifications is also needed. This description should generally include the following information:

a. Contents of the data set (variables, temporal frequency and range and spatial distribution)
b. Scientific rationale (intended uses, processing algorithms, quality control procedures, homogeneity, and discontinuities)
c. Scientific assessment (references, known deficiencies)
d. Ancillary information (volume, contacts, etc.)

To enhance integrated discovery of NOAA data and information, standards need to be applied in several areas. These include the following:

1) Discovery-level metadata content standard
2) Discovery-level metadata representation/exchange standard (syntax)
3) Discovery-level keyword vocabulary or lexicon (semantics)
4) Catalog search protocol specification

The requirement for standards and protocols to assist users in data discovery has been known for many years. Several initiatives to address this issue have been undertaken over the past decade and effective specifications have been developed.

There is no need for extensive evaluation of alternative discovery-level metadata content standards, since CSDGM is widely used and mandated by executive order. FGDC has agreed that the next version of the CSDGM will be in the form of a profile of ISO 19115 and GEOSS will likely adopt ISO 19115 as a metadata standard. Thus, compliance with ISO 19115 would be a beneficial side effect of conforming to CSDGM. It is recommended that CSDGM and ISO 19115 metadata standards be submitted for evaluation as possible NOAA standards (see Chapter 7).

The definition of keywords and a lexicon of terms that could be applied across NOAA is one of the most daunting short-term tasks required to improve integration of NOAA data management systems. Since NOAA includes components that span many different, albeit related, scientific fields, there is no external scientific/professional group that could serve as a forum for agreement on common terms Although the GCMD keywords for describing Earth science data and the World Meteorological Organization (WMO) keywords for describing meteorological and hydrological datasets provide a valuable starting point, neither is likely to be sufficient and neither include definitions for these keywords, a definite requirement for consistent, cross-NOAA implementation. Therefore, evaluating existing keywords and expanding these standard sets when necessary is a high priority.

The provision of services to discover data and products has traditionally been a serious weakness with operational systems. These systems rarely provide catalog services beyond a listing of products by, often arcane, file or product names. Instead, operational users have been expected to determine which products meet their needs through informal contacts with colleagues and staff at operational centers. To ensure operational products are fully utilized, it is essential that all NOAA products be fully described in catalogs that conform to accepted standards.

6.1.3. Extraction

After a user has determined that a data set or product is of value, they must request that the data be selected or extracted from the source file or database. If necessary, this step would include an identification step to determine if the user is authorized to access these data.

The extraction of data from data collections has often been a weak area in responding to user needs, primarily because of the multiplicity of formats presently produced by NOAA production systems. Frequently, NOAA data centers only provide the capability to select and receive entire files from collections that are organized into hundreds or thousands of such files. Several NOAA centers offer the capability to create files on demand by extracting data from databases or selecting subsets from existing collections. While some users might prefer to have data from several files aggregated into one, this capability is rarely offered.

Interoperability for extraction of NOAA data and information requires that data suppliers provide certain services for users. Users should be able to:

- Specify selection of sets (or subsets) of data (by time, area or parameters) without knowledge of or regard to the physical organization of these data into files.
- Perform standard, server-side processing to transform or reduce the volume of the data (*e.g.* averaging or sub-sampling over time or space).
- Specify how they would like to data organized for delivery (as a single file or as multiple files based on user-specified criteria).

Server-side Web services should be provided to support these capabilities.

Note that when these services modify the data, the metadata must be updated to correctly describe the data as it is delivered, rather than how it is archived.

6.1.4. Translation

There are many functions included within end-to-end data management. It is clear that developing and maintaining the tools required to support these functions represent a significant investment of resources. Many of these tools depend critically on the format of the data and associated metadata. The amount of software that needs to be created and maintained, and the resources required to accomplish this, therefore, depend critically on the number of formats used for similar or identical data types. Developing and maintaining multiple tools for visualizing multi-dimensional grids, for example, is not an effective application of resources. Integrating data management across NOAA requires decreasing the number of formats being used for particular data types and increasing the generality of software tools.

Although NOAA and its users will benefit from evolution to fewer, more general formats and improved tools, in general, it is not possible to maintain data or products in a form that exactly meet the requirements of all users. This is especially true for users outside of the data's traditional community or market. Thus, the utility of the information is greatly enhanced if it can be translated to meet the needs of different users. This may involve conversion of units, registration to a different reference system (*e.g.* political boundary to latitude/longitude) or map projection and translation of formats.

NOAA should review data representation (format) standards currently used for environmental data. It should agree on a small number of the most widely used formats to deliver digital NOAA data and products to its customers. It is likely that different standards will be most suitable for different types of data. Likewise, a format intended for interpretation by a computer will probably not be appropriate for interpretation by a person, and *vice versa.* Thus, users must be given the option to select (from a short list) the format that best meets their requirements.

Table 6.1 below is provided to show how format standards might be specified. Note that this is intended for illustration only since specific recommendations on standards will be determined through the standards process defined in Chapter 7.

Table 6.1

Data/product type	**Formats**	
	For interpretation by people	**For use by computers**
Publications	PDF, HTML	
Text products	ASCII, HTML, PDF	
Tabular data	PDF, HTML	Comma delimited ASCII, XML
Charts, graphs, maps	PDF, GIF, JPEG, PNG	BUFR, GML
Images (satellite, radar)	JPEG, GIF	BUFR, GML, GeoTIFF
Animations, image loops	GIF, MPEG, MOV, JPEG via Java applets	
Point/station data, soundings/profiles	PDF, HTML	Comma delimited ASCII, XML, netCDF, HDF5, BUFR
Time series data	PDF, HTML	Comma delimited ASCII, XML, netCDF, HDF5, BUFR
Multi-dimensional grids, large arrays	(see Charts, graphs, maps)	netCDF, GRIB, HDF4, HDF5, GeoTIFF

6.1.5. Delivery

Once the information (data, products and/or metadata) is in a form that can be used and understood by the user, it must be delivered in some fashion. It can be assembled as a file and transferred over a network or sent on physical media. It could also be accessed through an Application Program Interface (API), either as a file or as individual granules of data.

Delivery of data or information is ultimately the most fundamental procedure required to meet user needs. Since NOAA users span a wide range of capabilities, NOAA should provide a range of delivery options. This should include delivery via:

- Common Internet file transfer protocols (HTTP, FTP).
- Application Program Interfaces (APIs) following a client/server model (*e.g.* OPeNDAP).
- Web service interfaces (*e.g.* OGC interfaces such as WMS, WFS, WCS).
- Traditional postal delivery of digital media (CD-ROM and DVD) and hardcopy.

6.1.6. Application

It is assumed that technically advanced users would prefer to use their own applications (*e.g.* GIS, GrADS, Ferret, Integrated Data Viewer, etc.) to manipulate and visualize the data and products they receive. These users are best served by providing flexible and powerful extraction and delivery mechanisms that can provide data in a form recognized and compatible with commonly used visualization and analysis packages for scientific analysis.

The general public, normally having access to only a Web browser, prefers that NOAA perform the application function for them. For these users, NOAA should provide the capability to generate basic graphs (line and bar charts) and maps (contour, station plots) on the fly and deliver them as images (along the lines of Live Access Server or NCDC Climate Visualization applications).

6.2. Web Services

A Web service architecture that enables interoperable data access is often described in terms of a layered protocol stack, which includes:

- Transport layer, including protocols such as HTTP, FTP and others.
- Encoding layer, which assures that transported data are understood at either end of the transport.
- Service description, which describes the public interface for the Web service.
- Service discovery, which provides for easy publish/find functionality.

The widely-adopted W3C Web service specifications employ XML technologies to implement the encoding, description and discovery layers. The XML descriptions utilize metadata to describe both the semantic and syntactic aspects of the data being accessed, and these aspects provide a framework for distinguishing between the various technologies and implementations in common use.

Standards needed for transport and encoding are described above. Additional standards for services are needed if Web services are to fulfill their role within the implementation of GEO-IDE.

The NOAA GEO-IDE (through DMIT and the Data Management Information Architect) must define NOAA's core services necessary for both research and operations to fulfill NOAA's goals. These services must be built and prototyped in a step-by-step manner building on successful implementations. Service-level agreements specify acceptable uptime and reliability in measurable terms. Fundamental building blocks for good Service-level agreements include:

- **Defining exactly what the service to be provided is and by whom.**
  Specific NOAA Programs may have individual services; however they must all conform to standard interfaces depending upon the specific service, or level of service. Development of services must be coordinated by the data management information architect.

- **How the quality of service will be measured.**

  This includes the various levels of services that will be developed based on user requirements (*e.g.*, availability, throughput, latency, security, archiving, etc.).

- **How these performance levels be will reported and coordinated.**

  Component-level services such as modeling are institution-specific and under an SOA and Web services framework, that are outside the scope of responsibility of the framework. The interfaces between the various services or components require real-time monitoring and periodic adjustment for operations or as service requirements change with time.

- **Corrective actions the provider will take if service levels are not met?**

  NOAA-level management activities, as discussed in Governance and Road Map, require constant user feedback, monitoring and development through an iterative spiral development process.

A collection of initial services to be built will define NOAA's SOA. It is recommended that these key Services include (but not be limited to):

I. **Authentication Service**: Security services must be part of each service or application that participates in the SOA. An authentication service will identify and authenticate users to GEO-IDE and authorize access to data (some data is proprietary), determine quality of service levels (*e.g.* operational access to high-priority fast networks may be required), and provide reliable access to critical systems. This service will also provide single sign-on access to, and provide a common security infrastructure for the NOAA GEO-IDE.

II. **Registration Service**: This service will provide a common location where information about registered Web services is made available under GEO-IDE. It will contain metadata describing each service including availability, access restrictions, service-level agreements and service state.

III. **Data Cataloging Services**: These services will reside at multiple sites and provide a detailed listing of data holdings. Metadata will describe data content, restrictions, access methods, availability, etc. to be used by applications and other Web services. The catalog will also define how data is stored including available access methods (*e.g.* HTTP, FTP, gridFTP, OPeNDAP, etc.). Both public and private cataloging services may be available to serve general or restricted user communities.

IV. **Data Search Services**: These services will have knowledge of registered data catalog services and provide robust capabilities to locate GEO-IDE data holdings using standard search tools, ontologies, and metadata catalogs

V. **Data Delivery Services:** These services will provide format conversion and sub-setting, sub-sampling capabilities prior to delivery when requested. When data is available at multiple sites, these services will determine the optimal location and method by which to deliver the data. For example, it may be faster to obtain on-line data from a remote site, than to obtain it via the local mass storage system.

VI. **Notification Services**: These services will utilize Web service messaging to notify applications and other services when data products are available. They will be useful primarily for real-time operational environments where timely access to observational and model data is most critical.

VII. **Subscription Services**: These services will primarily be used to support real-time operational access to data via data push technologies and will complement access standard datasets including satellite data from Satellite Broadcast Networks. The services will permit applications or clients to automatically obtain data (via a notification service), once they are available. For example, forecasters at a weather forecast office can pre-stage the data they need before their shift begins using a subscription service.

VIII. **Real-time Monitoring Services**: These services will be used to monitor the state of NOAA's data systems and services under GEO-IDE. Monitoring will include metrics such as availability, redundancy levels, system loads, data volume, allowing additional resources or services to be added when required. For example, during hurricane season, additional resources could be required at the National Hurricane Center to handle data requests.

IX. **Workflow and Local Management Services**: These services will be used to manage application workflows in support of operational and time critical processes. Workflow management includes access to necessary data as well as computer and network resources required to run task and data dependent applications. These services will provide high levels of reliability and redundancy for workflows when required.

X. **Application Services**: These services will provide Web service interfaces to existing NOAA data management systems (*e.g.* CLASS, NOMADS, etc.) in order to provide generalized access for applications in the SOA. It will allow existing systems to be integrated into the SOA framework. For example, Web services will provide a means to discover and access data contained within the data management systems including cataloging, access restrictions, format conversion, and data update capabilities.

These services should be built based on increasing levels of complexity to allow for a learning curve within the organization and building on success.

In addition, user-based tools including portals and client applications must be developed or modified so they are Web-service aware and able to communicate with the services and data management systems available within GEO-IDE. These portals and tools must be built to satisfy diverse requirements (*e.g.* data access and discovery, system monitoring, model development, and verification), and diverse communities within and outside of NOAA.

7. The NOAA GEO-IDE Standards Process

7.1. Background

The lack of broad, uniform utilization of information technology (IT) standards that adequately meet the data integration needs of the agency is arguably the most acute factor contributing to the weakness of data integration within NOAA today. The explanations for this situation are both sociological – the lack of a tradition of close adherence to broad data standards – and technical – the limitations of the standards, themselves. The IT standards in existence today lack both the scope (range of applicability) and the depth (level of completeness in detail) required to adequately address NOAA's data interoperability needs. While the formulation of information management standards will inevitably lag behind the needs that drive them, the gap between formulation of standards and needs for the classes of data that NOAA holds is unacceptably long.

Since IT will likely continue its rapid evolution, the standards that support NOAA's data systems will also have to evolve. This is not an issue that can be addressed alone by NOAA since with the development of GEOSS, NOAA is, a partner within a larger data integration task force.

7.2. General Principles for the Standards Process

There are several general characteristics that are essential to an effective standards process:

- The process must be efficient – as streamlined as possible given other requirements.
- The process must be dynamic – standards will be updated or retired as technology best practices evolve. Standards that prove to have fundamental flaws should be quickly rejected.
- The process must be "open" – all ideas and approaches are on the table. There should be clear means for all interested parties to participate in the development of the standards and/or provide review input. The standard must be published and readily accessible to all interested parties.
- The process must be coordinated with other organizations that are facing related standards issues: Federal Agencies, National Forum for Geospatial Information Technology, GEOSS, WMO, IOOS, Global Climate Observing System (GCOS), Global Ocean Observing System (GOOS), Global Terrestrial Observing System (GTOS), etc.
- The process must be above conflict of interest – decisions should be made on the technical merits and cost/benefit considerations of the proposed ideas.
- The process must be methodical and evolutionary – harmonization of new standards with existing (successful) standards is essential.
- The process is inherently "layered" – Where broadly accepted industry standards exist, it is essential that these standards be adopted and used. But for these standards to be useful to NOAA, it is also essential that discipline-specific profiles, schemas, protocols and vocabularies be developed. For example, the industry standard SOAP protocol becomes an effective foundation standard for

NOAA only when it is intelligently augmented by standards that address 4-dimensional geospatial coordinates, useful structures built upon those coordinates (grids, time series, polygonal regions, etc.), and standards to represent discipline-specific variables (meteorology, ecosystems, etc.), units, measurement protocols, quality control, etc.

- The process must be well grounded in software practice. Ideas must be tested in functioning software before they can be regarded as accepted. Reference implementations of the standards should be encouraged wherever possible.
- Experience has shown that issuance of a mandate to adopt a suitable standard is not sufficient to ensure its widespread adoption and correct utilization within NOAA. GEO-IDE must include an outreach and education component to assure that NOAA data managers receive adequate training and support in the utilization of standards.

7.3. Related standards processes

NASA's Earth System Enterprise (ESE) has done a systematic evaluation of the standards processes adopted by a number of the community's most prominent standards-generating organizations, including ISO TC211, OGC, W3C, Consultative Committee for Space Data Systems (CCSDS), GRID computing, FGDC, Internet Engineering Task Force (IETF), and Sun's Java Community Process. Through an examination of the strengths and weaknesses of these many processes, a recommended process has been put forward. See http://seeds.gsfc.nasa.gov/stdprocRpt1.htm. NOAA GEO-IDE should leverage the NASA Earth System Data Standards Working Group (ESDSWG): Strategic Evolution of ESE Data Systems as the starting point for its standards process.

The ESDSWG Draft Standards Process (version 1.11, January 23, 2003) evaluates standards in three phases to assess how workable the implementation was and the level of success of the standard in operation. A successful outcome at each step in the process results in the submitted standard advancing along the following path toward approval.

Submitted Standard → Proposed Standard → Draft Standard → ESE Standard

At each step of the process, greater impact is accorded the standard.

- Submitted – No standing.
- Proposed – Affirmed that the proposed standard is applicable to ESE data systems
- Draft – Working implementations of the standard have been demonstrated in systems applicable to the ESE. ESE-funded data systems activities should consider use of this standard where applicable.
- Standard – Significant operational experience has demonstrated the value in ESE systems. Where applicable, ESE-funded data systems activities should use this standard or else justify why not. Use of this standard may be a requirement for future data systems awards.

7.4. Process for adoption of NOAA GEO-IDE standards

Similar to the ESDSWG, the focus within NOAA should be to review and evaluate existing standards and protocols that can contribute to further development of its data management systems. Where relevant industry standards exist, they should be given special consideration. Where useful standards exist, but they are found to be inadequate in scope, NOAA's preferred approach should be to participate in the standards process that created the standard in order to effect the changes needed.

Integration is enhanced by innovation and flexibility and an effective standards process must support these objectives. The proposed standards process is not intended to stipulate that a given standard CANNOT be utilized in a NOAA data system. Rather, the goal of this process is to identify standards that *must* be supported by information management systems (but not to the exclusion of alternative standards) and under exactly what circumstances they must be supported. Those circumstances may include such conditions as:

- NOAA systems must support either standard A or B.
- NOAA systems designated as 'operational' must support this standard.
- NOAA systems exchanging data as part of a service-oriented architecture must support this standard
- NOAA systems must be able to <u>ingest</u> data using this standard.
- NOAA systems must be able to <u>output</u> data in this standard

NOAA environmental information management systems exist or are being developed within a context of related efforts by other organizations to provide data standards, portals, catalogs, and gateways to environmental data and information resources. It is essential that NOAA information management systems be interoperable, both internal to the Agency and with respect to the broader geo-sciences community. Since the effectiveness of data standards to promote interoperability depends upon the chosen standards being few in number and widely adopted, the philosophy regarding the selection of standards within NOAA is first to adopt, then to adapt, and only as a last resort to create standards. New standards should be developed only as a last resort when it is proven that no existing standards are, or could be, applicable. In those exceptional cases where it is agreed that a new standard is needed, that standard should be developed in close partnership with the appropriate community of domain specialists outside of NOAA as well as inside. The process for development of new standards should lead to their eventual certification by a widely recognized standards-generating organization.

NOAA must address both short-term and long-term standards challenges. The short-term challenge is to adopt an initial collection of standards, as rapidly as feasible, sufficient to begin to glue NOAA's many separate data systems into an interoperable framework. The long-term challenge is to achieve a steady state in which new standards are adopted and old standards are phased out in keeping with the evolution of technology and needs. The process defined below describes the long-term, steady-state procedure for adoption and review of standards. A fast-track process for defining an initial set of standards is proposed at the end of this chapter.

As shown below, NOAA standards will progress through three phases of evaluation and adoption before being accepted as a NOAA Standard. At each step of the process, greater importance would be accorded the standard within NOAA.

Submitted Standard ➔ Proposed NOAA Standard ➔ Recommended NOAA Standard ➔ NOAA Standard

In essence the phases would be:

1. **Submitted Standard** – Evaluation of the requirement
   a. Is there, or will there soon, be a need for this standard?
2. **Proposed NOAA Standard** – Technical evaluation and request for comments
   a. Is the standard technically sound?
   b. Have working implementations of the standard been demonstrated in environmental information systems within NOAA?
   c. Does it measure up well compared with alternatives?
   d. Is it well understood and well documented?
3. **Recommended NOAA Standard** – Evaluation in real-world NOAA systems
   a. Resources may be dedicated to define extensions to the standard that are needed to meet NOAA requirements and to develop additional software tools needed to support and simplify implementation. These activities will be coordinated with NOAA partners (*e.g.* US-GEO, GEOSS, IOOS) when appropriate.
   b. After significant experience with the standard has been gained in environmental information systems within NOAA, determine if the standard meets the requirements for which it was proposed.
4. **NOAA Standard** – Approved and mandated where appropriate.

The detailed actions and responsibilities of the individuals and groups contributing to the proposed standards process within NOAA are described in Table 7.1 below.

Table 7.1

Submitted Standard – At this level, a standard has no standing within NOAA.	
Submission	A standard can be submitted by any NOAA employee for consideration as a NOAA standard. The request should be submitted to DMIT through the responsible DMIT LO or Goal Representative. The request should include at least the following information: a. Standard name. b. Authority responsible for the standard (If the submitter is proposing a new standard that has been developed within NOAA, he/she must provide justification describing why no existing standard is applicable or sufficient). c. If applicable, statutory requirements for supporting the standard (Executive order, international agreement, etc.). d. Significant applications (within and outside NOAA) currently using the standard. e. Purpose/application (File transfer, API, delivery format, web services, etc.). f. Proposed data type(s) to which the standard would apply.

Evaluation	DMIT will evaluate the standard to determine if it addresses a real need and if it could be applicable to NOAA environmental information systems (*i.e.* does it address a need for a standard relevant to NOAA, could it apply to NOAA data, etc.). The DMIT review will be concerned with the possible requirement for and utility of such a standard within NOAA and will <u>not</u> consider its technical merits. DMIT should complete its evaluation within 60 days of the submission.
Conclusion	DMIT will respond to the submitter with the results of its evaluation and the reasons for its conclusion. If the standard is rejected, the submitter will be given an opportunity to respond to the particular issues responsible for the rejection. If the standard is accepted, DMIT will request additional information from the submitter (described below) and the standard advances to the next stage of the process.
Proposed NOAA Standard – At this level, the standard has status within NOAA and can be provisionally utilized by NOAA information systems for evaluation.	
Submission	DMIT will request the following additional information from the original submitter: a. Existing software tools that support the implementation of the standard. b. Detailed description/definition of the standard. Upon receiving this information, DMIT will compile all of the information pertaining to the standard and submit the standard as a Proposed Standard to a team or group designated by the DMC. If the DMC approves, the standard will be accorded the status of a Proposed Standard within NOAA and the standard and its supporting information will be entered into the NOAA Guide on Integrated Data Management.

Evaluation	The GEO-IDE Project Office (or the NOAA Data Management Integration Architect (DMIA)) will advertise the standard throughout NOAA. System managers and developers will be asked to consider the applicability of the standard to their systems. Where applicable, they will be encouraged to examine the technical feasibility of implementing the standard. This advertisement and evaluation would be considered a Request for Comments (RFC) on the standard. The GEO-IDE Project Office/DMIA will: a. Coordinate, and may provide support for development of any extensions that are needed to meet NOAA requirements. b. Coordinate and support development of pilot software tools to simplify implementation of the standard. c. Support pilot projects that apply the standard, with priority for projects that demonstrate integration of data across line office or program boundaries. d. Coordinate application of the standard with relevant NOAA partners (*e.g.* NASA, IOOS, US-GEO, etc.). System managers should report to the DMIA the response of implementers and users in the application of the standard and how the standard interfaces and operates with related standards. The DMIA will in turn pass this information to DMIT for its consideration.
Conclusion	DMIT will periodically review the evaluation reports and comments it has received pertaining to each Proposed Standard. It will determine if: a. Working implementations of the standard have been demonstrated in environmental information systems within NOAA. b. Use of the standard has been shown to contribute to the NOAA mission. c. Users report if the use of the standard has a positive impact on their interactions with NOAA. If the answers to the above questions are positive, the standard will advance to the next stage of the process. If the answers to the above questions are negative, DMIT can recommend to the DMC that the standard continue to be considered a Proposed Standard or that it be rejected and its entry marked as such in the NOAA Guide on Integrated Data Management.

Recommended NOAA Standard – At this level, all NOAA data systems should consider supporting the standard wherever applicable.	
Submission	DMIT submits the standard to the team or group designated by the DMC for consideration as a Recommended Standard. If the DMC approves, the standard will be accorded the status as a Recommended Standard within NOAA. The standard's entry in the NOAA Guide on Integrated Data Management will be modified to reflect its status as a Recommended Standard.
Evaluation	Starting with managers and developers identified during the evaluation of the standard as a Proposed Standard, the GEO-IDE Project Office/DMIA will: a. Identify in-house experts to develop training material and to assist developers in applying the standard. b. Identify any additional software tools that are available to support and simplify implementation. c. Support the definition of any additional extensions needed to meet NOAA requirements and coordinate this work with relevant partner agencies and organizations. d. Coordinate and support development of operationally robust software tools to aid in implementation of the standard. Where the standard could be applicable to NOAA partners (US-GEO, GEOSS, IOOS, etc.) these activities will be coordinated and pursued in cooperation with the relevant partners. The entry for the standard within the NOAA Guide on Integrated Data Management will be updated to reflect the availability of additional material supporting its implementation (expertise, training, and software) as it becomes available. Managers and developers of all NOAA environmental information systems should consider supporting the standard wherever applicable. System managers should report their experience with the standard to the Project Office.
Conclusion	DMIT will periodically review the status of implementation of each Recommended Standard. DMIT will determine whether: a. The standard is sufficient to meet NOAA requirements. b. Satisfactory training material and software tools are available to support implementation of the standard. c. Significant operational experience with the standard has been gained in environmental information systems within NOAA. d. The value of the standard in NOAA environmental information systems has been demonstrated. If the answers to the above questions are positive, DMIT will submit the standard to the DMC for consideration as an Approved Standard. If the answers the above questions are negative, DMIT can recommend to the DMC that:

	a. The standard continue to be considered a Recommended Standard. b. The standard be demoted for further evaluation as a Proposed Standard. c. The standard be rejected and its entry marked as such in the NOAA Guide on Integrated Data Management.
NOAA Standard – NOAA data systems should support this standard wherever applicable.	
Submission	DMIT submits the standard to the DMC for consideration as an Approved Standard. If the DMC concurs, it in turn submits the standard to the NOSC for its consideration. If appropriate, the standard will be submitted to wider outside groups (US-GEO, GEOSS) for their endorsement and their response will be considered by the NOSC in its deliberations. If approved by the NOSC, the standard will be accorded the status as a NOAA Standard and its entry in the NOAA Guide on Integrated Data Management will be modified to reflect its upgraded status.
On-going use and evaluation	NOAA data systems should support this standard wherever applicable or else justify why not. Compliance with this standard may be a requirement when considering funding for future data systems. DMIT will periodically review the status of implementation of Approved Standards to ensure they are supported wherever applicable. If a standard falls out of use, becomes obsolete, or is superseded by another standard, DMIT may recommend to the DMC (and consequently the NOSC) that the standard be marked as deprecated within the NOAA Guide on Integrated Data Management. After being marked as deprecated for 2 years, its entry may be removed from the Guide.

7.5. Proposed process for defining an initial set of standards

Since many standards currently in use within the geosciences community could be applicable to NOAA, an alternative process is recommended for fast-track submission of an initial set of standards. It is proposed that a NOAA standards workshop be held to:

a. Approve the standards process itself (using this document as a straw man proposal).
b. Evaluate an initial set of standards for consideration as NOAA standards (a list of proposed standards will be prepared by the DMIT Standards Sub-group and circulated in advance of the workshop).
c. Recommend standards that should immediately be submitted to the DMC as Proposed NOAA Standards.

8. NOAA Guide on Integrated Information Management

GEO-IDE provides a framework to guide development of data management systems within NOAA and includes general guidance applicable to all environmental data and information management systems. As noted in Chapter 2, GEO-IDE does not attempt to stipulate or define all of the details necessary to properly plan, implement, and execute the data management components of any particular observation or data processing program within NOAA. This document provides an overall framework in which detailed plans for these programs should be defined. To assist program and project managers, DMIT recommends that a NOAA Guide on Integrated Information Management be developed. The Guide is intended to help NOAA program and project managers find resources needed to help write detailed implementation plans for their programs that conform to GEO-IDE recommendations. These resources include NOAA data management polices and guidelines, an inventory of data systems, relevant NOAA, national and international standards, and recommendations for developing data management plans.

Due to the dynamic nature of data management and the technology that supports it, the NOAA Guide on Integrated Information Management will be published and maintained as an on-line document. The guide will be maintained online at http://www.cio.noaa.gov/dm/geoide by joint cooperation between the NOAA CIO Policy Office and members of DMIT, at least until the NOAA Data Management Information Architect/DMIT Chair is hired.

Guide components

The guide will contain the following sections:

1. NOAA and other Federally-mandated data management policies.
2. NOAA-wide standards (in all stages of approval).
3. Registry of information management data sources, tools.
4. Guidelines and/or checklist for creating and implementing NOAA project-level data management implementation plans

The following discussion provides details about each section.

8.1. Data management policies

New and existing data management systems should support data management policies established by the Office of Management and Budget (OMB), the National Archives and Records Administration (NARA), as well as those defined by the Department of Commerce (DOC) and NOAA. We use ‘data management policy’ in this context to mean rules or tasks defined in legislation or by a regulatory group that define a required or recommended action to be performed by a data manager.

The guide will provide a description of and pointers to NOAA data management policy documents. Because the emphasis of the guide is on actions that NOAA data managers are expected to perform, it will provide more details about NOAA policies. The guide will also provide brief descriptions and links to other federal data management policy documents, including those from organizations such as OMB, NARA, FGDC, and others.

As an example, all NOAA Websites are required to comply with DOC Web policies, listed at http://www.osec.doc.gov/webresources/DOCWebPolicies_BestPractices.html.

While it is not a requirement for all NOAA Websites to comply with DOC Best Practices (listed at the same Web page), NOAA should follow DOC Best Practices whenever practical. An entry in the guide might include a brief definition of the DOC policy about Website contact information and a link to the DOC-maintained page defining that policy (see Example 1). Alternatively, the guide could describe the policy framework and reference the detailed resource (Example 2).

Example 1. Example of a brief entry describing a specific data management policy statement.

> **Policy**: Every Website of a Department of Commerce organization shall provide an electronic method for comments, inquiries and accessibility issues.
> Link : http://www.osec.doc.gov/webresources/Policy4_ContactInfo.html
> Maintaining Organization: Department of Commerce.

Example 2. Example of an entry describing a collection of data management policies.

> **1. Policy:**
> Compliance with the Department of Commerce Web Policies as listed at http://www.osec.doc.gov/webresources/ is mandatory for all NOAA Websites. Department of Commerce Best Practices are not mandatory, but should be followed where feasible and practical.
>
> **2. Purpose and Authority:**
> The purpose of this policy is to ensure that NOAA Websites comply with DOC Web policies, which have been developed to ensure compliance with all applicable laws and government directives.
>
> **3. Scope:**
> This policy applies to all NOAA Websites, public and internal.
>
> **4. Terms and Discussion:**
> NOAA Web pages have a high degree of visibility and represent the official position of NOAA. It is imperative that there is consistency and quality within NOAA's Websites regarding identification, privacy, accessibility, and usability. Please see each DOC Web policy for its effective date.

Some data management policies are defined in NOAA Administrative Orders (Example 3) or other official NOAA communications and apply to all NOAA components. These NOAA policies should be organized together in the guide, with links to the appropriate NOAA resources and with references to related resources from OMB, NARA, or other federal entities.

Example 3. Example of guide description of a NOAA Administrative Order directive relating to data management.

NOAA Administrative Order 216-101, Ocean Data Acquisition. This NAO establishes policies and procedures to ensure that NOAA ocean data collection activities including open-ocean, Great Lakes, coastal, and estuarine data collection activities support multiple uses of those data for purposes other than those for which they were originally collected." (http://www.rdc.noaa.gov/~nao/216-101.html). Additional details about how to interpret this NAO could also be included in the guide if resources are available to do so.

Data management policies and directives that apply to NOAA data management practices may also be gleaned from the following sites (and many others):
OMB example: http://www.whitehouse.gov/omb/egov/b-1-information.html

OMB example: http://www.whitehouse.gov/omb/egov/documents/fea-drm1.PDF
DOC example: http://www.osec.doc.gov/webresources/
NOAA example: https://secure.cio.noaa.gov/hpcc/docita/
NOAA example: http://www.cio.noaa.gov/itmanagement/ppaochg_index.html
NOAA example: http://www.cio.noaa.gov/itmanagement/ciopol.htm
Other example: http://cio.gov

8.2. NOAA-wide standards

NOAA data management efforts need to incorporate widely accepted international and national standards in order to share data throughout the agency and with its many external customers. Common naming standards and definitions, location standards and syntax must be adopted in order to gain the maximum value.

The guide will serve as a collection point and resource to identify applicable NOAA standards. It is anticipated that NOAA will adhere to national and international data management standards such as the FGDC Content Standard for Digital Geospatial Metadata and/or the ISO 19115 and related metadata content standards. Chapter 7 discusses the procedures for identifying, evaluating and establishing NOAA data management standards in detail.

Some examples of standards that NOAA data management systems should follow include:

DOC Best Practice example: http://www.osec.doc.gov/webresources/BP5_XHTML.htm
NOAA CIO Standards example: https://secure.cio.noaa.gov/hpcc/noaaita/
Other Standards example: http://www.fgdc.gov/standards/standards.html.

8.3. Registry of data management software (applications and tools)

The guide should include an inventory or registry of supported NOAA data management products or tools that facilitate life cycle data management tasks. These tools should be made available widely to all potential users to encourage common use and to accelerate developing data discovery and delivery systems. Maintaining an inventory of data management tools and products should expose duplications and inefficiencies, as well as identify opportunities for partnering with ‘power users’ of specific tools. Ultimately, a unified portal that identifies all data products created and distributed by NOAA would greatly improve the ability of data managers and data users to find and use those products.

Examples of data management tool registries include:

The NOAA Observing Systems Architecture (NOSA) observing system inventory at http://www.nosa.noaa.gov/observing_systems.html

The Component Registration and Organization Environment (CORE) of the Federal Enterprise Architecture initiative at http://Core.gov and https://www.core.gov/reusecomponent.html. CORE will become a networked community of component developers and users, and will offer numerous components of various types and complexities, including business components, e-forms and technical components. Using the CollabNet SourceCast tool, CORE.GOV’s robust collaborative environment can organize and map components in a variety of ways to make them easy to identify, discuss, and develop.

The FGDC metadata tools inventory at http://www.fgdc.gov/metadata/metatool.html.

8.4. Data management planning template

A Data Management Planning Template includes the necessary steps that program and project managers must take to address the design, budgeting, and planning of information assets. With the implementation of the PPBES throughout NOAA, this plan identifies the strategies, activities and projects related to corporate data management goals. Moving to a shared-data, integrated information environment will take a well thought-out plan. Integration is not something that takes place across all systems at one time. Rather, it must proceed with one or a few projects at a time. Selection and prioritization of projects will be key to success on several fronts with the greatest performance outcomes.

Example of PPBES and IT Planning:

NOAA PPBES IT Plan 2005-2010 Example 1:
http://www.cio.noaa.gov/itmanagement/NOAAstrategicITplan_Architecture2005.pdf
NOAA example 2: http://www.nosa.noaa.gov/ppbes.html
NOAA PPBES PowerPoint. Example 3:
http://www.ppi.noaa.gov/PowerPoint/PPBES_Process.ppt

Other resources that should help project managers plan for data management include:

NOAA CIO Enterprise Architecture example: https://secure.cio.noaa.gov/hpcc/noaaita/
(to access use NOAA e-mail user login/password)
Other example from DMAC Plan: http://dmac.ocean.us/dacsc/imp_plan.jsp

A comprehensive data management plan should address each of the functional areas broadly outlined below. The guide should provide information about how to assure that each of these functional areas are addressed. It should provide the link to the appropriate policies and standards that apply to each of these functional areas, so that planners can feel reasonably assured that their plans comply with all identified NOAA policies, adhere to all NOAA standards, utilize approved or recommended data management tools whenever possible, and minimize duplication of application or data products.

Components of a comprehensive data management plan should address the following functional areas:

1. Interface to the Observing System:
    - Real-time, on-site – quality control procedures
    - Data transmission and collection mechanisms (to processing centers and direct to users)
    - Monitoring the observing system (a systems approach)
    - Providing for (future) optimal system design
    - Standards used for data format, semantics and transmission
2. Data Assembly ("Centers of Data"):
    - Additional real-time quality control procedures
    - Real-time product generation and distribution
    - The role of Centers of Data (e.g. the CLASS collection, NDBC, …)
    - Delayed mode/retrospective quality control
3. Archive:
    - Ensure that every data stream has a designated archive

- Creation of archive products (e.g. climate data records)
- Develop and implement archive and retention policies

4. Data and product discovery and on-line browse:
    - Metadata standards use (content, syntax, semantics)
    - Metadata management
    - Metadata searching (access protocols and interface with community dataset catalogs (NOAA Server, GCMD)
    - On-line visualization
5. Interoperable Data Access (Transport):
    - Push versus pull
    - Standards and protocols supported
    - Single standard, multiple standards, gateways
    - Extensibility
    - Special data types (*e.g.* video)
6. Operations:
    - Fault tolerance
    - Security
    - Continuity/robustness
7. Users, Information Products and Applications:
    - Determining requirements
    - Soliciting feedback
    - Who creates information products? Interaction with the private sector.
    - Real-time notification of availability of information products

9. Priorities for Action

This section summarizes the list of activities that NOAA should undertake to set this plan in motion and begin the journey of working towards an integrated data management environment within NOAA. The following activities have been identified for FY 2007-2010 as having the highest priority. These activities, some of which are underway, are seen as requiring immediate attention because they will have the highest payoff to NOAA in the near term. These key activities were selected because they are critical to NOAA's mission, benefit the end user, can be delivered in a fairly short time, or are essential to creating an integrated data management environment within NOAA.

1. Establish the GEO-IDE project management structure.
2. Secure funding to support GEO-IDE activities.
3. Identify major information management systems in NOAA.
4. Evaluate, adopt and adapt information management standards within NOAA and publicize them via an on-line NOAA Guide to Integrated Information Management.
5. Define a NOAA-wide, service-oriented Web architecture.
6. Test the feasibility of utilizing a "data typing" approach to NOAA data and refine categorization of data types used throughout NOAA.
7. Develop and acquire technical knowledge and skills.
8. Identify technologies for implementation of the SOA, define core Web services needed and implement these services via pilot projects.
9. Investigate new technologies to support the NOAA mission.

The critical factor in the success of any set of project initiative for NOAA is the need for dedicated leadership, staff and project teams within the agency to coordinate all activities over a period of time. As NOAA moves toward the goal of integrated data management, the challenge will be to continue to identify goals and activities, measure progress, work towards measurable outcomes and deliverables, communicate accomplishments, listen to critical feedback, learn from successes and failures, and continually reinvent ourselves, in light of constrained resources.

Appendices

Acronyms

ANSI	American National Standards Institute
API	Application Programmer Interface
AWIPS	Advanced Weather Information Processing System
BUFR	Binary Universal Form for the Representation (WMO)
CCSDS	Consultative Committee for Space Data Systems
CFO	Chief Financial Officer
CIO	Chief Information Officer
CLASS	Comprehensive Large Array-data Stewardship System
CORE	Component Registration and Organization Environment
CSC	Coastal Services Center
CSDGM	Content Standard for Digital Geospatial Metadata
DMAC	Data Management and Communications (component of IOOS)
DMC	Data Management Committee
DMIT	Data Management Integration Team
DOC	Department of Commerce
EOSDIS	Earth Observing System Data and Information System
ESDSWG	Earth System Data Standards Working Group (NASA)
ESE	Earth System Enterprise (NASA)
ESRL	Earth Systems Research Laboratory
FGDC	Federal Geographic Data Committee
FTP	File Transfer Protocol
F/NWC	Fisheries/Northwest Center
F/SWC	Fisheries/Southwest Center
GCMD	Global Change Master Directory
GCOS	Global Climate Observing System
GEO-IDE	Global Earth Observation – Integrated Data Environment
GEOSS	Global Earth Observation System of Systems
GeoTIFF	Geographic TIFF
GIF	Graphics Interchange Format
GIS	Geographic Information System
GML	Geographic Markup Language
GOOS	Global Ocean Observing System
GrADS	Grid Analysis and Display System
GRIB	Gridded in Binary (WMO)
GTOS	Global Terrestrial Observing System
HDF	Hierarchical Data Format
HTML	Hypertext Markup Language
HTTP	Hypertext Transport Protocol
IEOS	Integrated Earth Observation System
IETF	Internet Engineering Task Force
OMB	Office of Management and Budget
IOOS	Integrated Ocean Observing System
ISO	International Organization for Standardization
IT	Information Technology

IWGEO	Interagency Working-group for Global Earth Observations
JPEG	Joint Photographic Experts Group
LAS	Live Access Server
MADIS	Meteorological Assimilation Data Ingest System
MPEG	Moving Picture Experts Group
NAO	NOAA Administrative Order
NARA	National Archives and Records Administration
NASA	National Aeronautics and Space Administration
NCDC	National Climatic Data Center
NCDDC	National Coastal Data Development Center
NCEP	National Centers for Environmental Prediction
NDBC	National Data Buoy Center
NESDIS	National Environmental Satellite, Data and Information Service
netCDF	Network Common Data Form
NGDC	National Geophysical Data Center
NMFS	National Marine Fisheries Service
NOAA	National Oceanic and Atmospheric Administration
NOMADS	NOAA Operational Model Archive and Distribution System
NOS	National Ocean Service
NOSA	NOAA Observing Systems Architecture
NOSC	NOAA Observing System Council
NSDI	National Spatial Data Infrastructure
NWS	National Weather Service
POP	Program Operating Plan
PDF	Portable Document Format
PMEL	Pacific Marine Environment Laboratory
PNG	Portable Network Graphics
PPBES	Planning, Programming, Budgeting, and. Execution System
OAR	Office of Oceanic and Atmospheric Research
OGC	Open Geospatial Consortium
OPeNDAP	Open source Project for a Network Data Access Protocol
SOA	Service-Oriented Architecture
THREDDS	Thematic Real-time Environmental Distributed Data Services
TIFF	Tagged Image File Format
UCAR	University Corporation for Atmospheric Research
US-GEO	US-Global Earth Observation System
WMO	World Meteorological Organization
WSDL	Web Services Definition Language
W3C	World Wide Web Consortium
XML	Extensible Markup Language

Membership of NOAA DMIT

Jordan Alpert, NWS NCEP
Landry Bernard*, NWS NDBC
Sarah Brabson, Office of the CIO
Tina Chang, NMFS Headquarters
Don Collins, NESDIS Headquarters
Joseph Facundo, NWS Observing Systems Branch
Mark Govett, OAR ESRL
Ted Habermann, NESDIS NGDC
Steve Hankin*, OAR PMEL (co-chair)
Andrea Hardy, NOS IOOS
Richard Kang, NMFS F/NWC
Tony Lavoi*, NOS CSC
David Layton, NESDIS Headquarters
David McGuirk, NESDIS NCDC Consultant
Roy Mendelssohn*, NMFS F/SWC6
Russ Rew*, UCAR Unidata
Glenn Rutledge, NESDIS NCDC (co-chair)
Susan Starke, NESDIS NCDDC
David Stein, NOS CSC

* - Also member of the DMAC Steering Team